# nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

## THE CIRCULAR ECONOMY

A world without waste

PAGE 435

MARINE BIOLOGY

## CAUGHT IN THE FOOD WEB

*There's more to jellyfish than meets the eye*

PAGE 432

PLANETARY SCIENCE

## THE MOON IN A SPIN

*Polar wander signals ancient origin for lunar water*

PAGES 455 & 480

BEHAVIOURAL ECONOMICS

## TOP-DOWN MORALITY

*Political corruption breeds individual dishonesty*

PAGES 456 & 496

NATURE.COM/NATURE

24 March 2016 £10

Vol 531, No. 7595

# THIS WEEK

## EDITORIALS

# Cultural conundrum

*The Chinese government's professed commitment to transparency and responsiveness has had a rocky start. That bodes ill for the desire to attract the best science brains from around the world.*

The Chinese government is open and accountable. Says who? Says the Chinese government. In mid-February, the state council, the nation's highest administrative authority, released a statement saying that government affairs should be "transparent to, understood by, and responsive to the public".

Ten days later, it announced a new social-media app to allow interaction with the government and distribute information about government services, boasting the slogan "the government is right by your side".

And earlier this month, premier Li Keqiang told all ministers that they would have to make themselves available to the press. Sure enough, at the end of the annual meetings of the People's Congress and the Chinese People's Political Consultative Conference (CPCC), a couple of dozen ministers appeared in the Chinese media to discuss the meetings' proceedings and the future of the country.

But China's version of transparency is very one-sided. Information flows freely in one direction, and it is not towards the people. For Chinese people even to ask questions of the government remains a no-no.

President Xi Jinping made this clear during a tour of China's three most prominent state-owned media outlets, when he encouraged journalists to toe the party line. His message was clear: the media are a propaganda tool of the state rather than an outlet for public discussion.

'Don't ask, we will tell', has also been the government's approach to its environmental problems. Although the government often laments its air and water pollution problems in the state media, it does not appreciate being questioned about how well it is dealing with them. Last year, it quickly blocked *Under the Dome*, a video documentary by journalist Chai Jing detailing the problem, which had gone viral.

The freedom to question is a hot topic in China right now. Last week, amid the government's latest responsiveness and transparency campaign, a journalist for the *Science and Technology Daily* asked a CPCC member an innocuous question on a topic discussed in the previous day's government meeting. Officials had mentioned that some industries that served the military had major down time between military projects, causing gaps in productivity. The journalist wanted to know what might be done to help these industries, and whether lessons from other countries might offer any clues. The CPCC member berated the journalist in front of the whole forum: "I'd like to point out that some media, for example, you, the one from *S&T Daily*, dwell on some negative issues," he said. "I've already noted down your licence number, so be careful, or you'll have to answer to the authorities." The journalist said that there must have been a misunderstanding and offered to discuss the matter later. The bureaucrat refused: "I don't have an obligation to talk with you," he said.

One of China's major goals is social harmony, and it remains an open question whether a democratic or an autocratic form of government is best able to achieve that. The Chinese state media have been quick to point to the current US presidential race as a clear sign of the 'malfunction' of democracy, because people 'vote to vent', not to choose a good leader. But China holds tightly to another goal as well — scientific development and innovation — for which free questioning and open debate are essential. Restricting the ability to ask questions does not work for that. It doesn't work for keeping scientific experiments on course, or for making sure that publications are as good as they can be. It doesn't work for confirming the details of a potential scientific collaboration, or for ensuring that grant committees select the most promising projects. It doesn't work for nailing down the details of material-transfer agreements, or for picking the best science-policy objectives. And it won't help China's ongoing efforts to lure the best brains from around the world. From the fear of increasingly strict regulations on what can be said to the ban on the use of tools such as Google and Google Scholar, which many scientists consider essential for information gathering, the country is making itself a much less attractive destination.

***"For Chinese people even to ask questions of the government remains a no-no."***

If China is to meet its scientific objectives (see page 424), a culture of debate and transparency is essential. No scientific community does this perfectly, but, in a country that discourages questions, the will to investigate and fully understand cannot be expected to take root. ■

# Siren call

*Now that gravitational waves have been discovered, it is time to put them to use.*

The Universe is big, and getting bigger all the time. A little-known fact about gravitational waves, the latest cosmological discovery to get physicists excited (see page 428) is that they could help to measure this expansion. And they could show why the expansion has been accelerating, rather than slowing down as expected, under the push of a mysterious force dubbed dark energy.

The way in which astronomers conventionally measure distances has ancient roots. With ingenuity and a dash of basic trigonometry, the ancient Greek astronomer Aristarchus of Samos was able to measure the Moon's distance from Earth with surprising accuracy — in the third century BC.

A similar method to Aristarchus', using a concept known as stellar parallax, was first applied to measure a star's distance from Earth in 1838, and is still used today. The European Space Agency's Gaia probe is currently using it to compile a state-of-the-art catalogue of one billion stars in the Milky Way, extending the reach of parallax to unprecedented distances and cutting errors down to less than 1%.

Stellar parallax is good, but it can go only so far. It entails measuring a star's apparent position in the sky at different times of the year, as Earth (or a space probe such as Gaia) orbits the Sun. The distance between the two observing points, measured to high accuracy, provides the base of a triangle. The distant star is at the opposite vertex. The smaller the angle at that vertex, the farther away the star is.

But because the size of Earth's orbit is fixed, as the vertex moves farther away the angle becomes smaller and smaller, and ultimately impossible to measure with any accuracy. (The basic unit of measurement of astronomical distance, the parsec, is short for 'parallax of one arcsecond', which refers to the size of that angle. One arcsecond is 1/3,600th of a degree, and in typical parallax measurements the angles are much smaller.)

For objects in more distant galaxies, astronomers have devised steps that build on the parallax method. Each step is a 'rung' on what they call the cosmic distance ladder. For example, the distance from Earth of the Andromeda Galaxy, the closest large galaxy to the Milky Way, is estimated by measuring the brightnesses of various types of star in it and comparing them to the brightnesses of similar stars closer to Earth whose parallax is known. Such estimates exploit the fact that similar stars look fainter the farther away they are.

Andromeda is roughly 780 kiloparsecs (2.54 million light years) away. Telescopes cannot resolve individual stars in galaxies that are hundreds of millions of parsecs away — except when those stars happen to blow up as supernovae. Astronomers use some supernovae as signposts of cosmic distances, or 'standard candles', meaning that their measured brightness is an indicator of their distance.

A major complicating factor is that the observed brightness of distant objects can be affected by foreground matter such as dust. Wouldn't it be wonderful to have a more direct and reliable way of measuring distances — one that were as precise as Gaia and worked at scales from the galactic to the cosmic?

Beginning with a paper in this journal 30 years ago (B. F. Schutz *Nature* **323,** 310–311; 1986), physicists have suggested that gravitational waves could provide such a tool. The ripples, predicted by Albert Einstein in 1916 as a consequence of his general theory of relativity, travel across the Universe without being dimmed significantly by dust or gas.

***"Stellar parallax is good, but it can go only so far."***

The gravitational waves that struck Earth in September and were recorded by the Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) carried information that revealed their strength at the source. In theory, this information can be used to work out the source's distance.

In the next few years, other interferometers are scheduled to join LIGO to form a global network of gravitational-wave observatories. Together, these instruments could calculate the positions and distances of merger events. Neutron-star mergers are especially interesting to cosmologists because they should also produce bursts of short, high-energy γ-rays, which would help to pinpoint their galaxies of origin.

Researchers hope that they will be able to use information from mergers as a way to calculate the distances of known galaxies. Because gravitational waves are more similar to sound than they are to light, physicists have dubbed these potential signposts 'standard sirens'.

One of the main uses of supernova standard candles has been to measure the current rate of cosmic expansion. Standard sirens could provide an independent way to do this. And, if space-based interferometers are added to the network, they could be used to track dark energy. Hear the call. ■

# Power of the pen

*Scientists must unite to stop Turkey from removing the right to freedom of expression.*

When he labelled outspoken academics as terrorists, Turkey's increasingly authoritarian President Recep Tayyip Erdoğan was probably not thinking of Voltaire's eighteenth-century philosophical maxim: "To hold a pen is to be at war".

Erdoğan sent shivers down the spines of those who care about human rights by declaring on 14 March that those who support terrorists are as guilty as those "who pull the trigger", and that Turkish law should be changed to reflect this. "The fact that an individual is a deputy, an academic, an author, a journalist or the director of an NGO does not change the fact that that person is a terrorist," he said.

One the same day, three academics from universities in Istanbul were hauled into police custody and then refused bail while prosecutors considered charges of making propaganda for a terrorist organization.

Their crime? In January, they had signed a petition that called for an end to violence in the southeast of the country, where government forces have been fighting Kurdish separatists. The petition was signed by 1,128 academics, mostly from Turkish universities, when it was publicly launched on 11 January. It immediately sparked Erdoğan's rage. Many politically appointed university rectors leapt into line, launching disciplinary investigations into members of their staff who had signed — more than 500 so far. Dozens of signatories were brought in for police questioning. The harsh response attracted a shocked solidarity. Another 1,000 people signed the petition, including a large number of Western scientists, before it was closed on 20 January.

An atmosphere of uncertainty and fear prevails. None of the signatories knows whether they, too, will be arrested, and several have had death threats. Some have actively sought sabbaticals abroad; those working outside the country are afraid to return even to visit family.

Meanwhile, Turkey is playing a major part on the world political stage, in a role that is overshadowing the fate of the academics.

Turkey is a geopolitical fulcrum. On one side it borders war-torn Middle East, on the other, strife-ridden Europe that is struggling to cope with the refugee crisis. When the country reached a historic agreement with the European Union last week to take back migrants who were crossing into Europe illegally, many in the EU complained bitterly about making a deal with Erdoğan because of his worrying human-rights record.

Terrorist attacks in Turkey are intensifying, some carried out by Kurdish separatists, others by the Islamist group ISIS. Erdoğan's controversial announcement followed on the heels of a deadly attack in Ankara, and on 19 March, a suicide bomber killed four in Istanbul. Kurdish separatist terrorism had abated during a two-year ceasefire, but that broke down last July. Erdoğan argues that the peace petition, by focusing only on government military attacks on Kurdish militants, which have killed many innocent civilians, and ignoring terrorist attacks and other serious human-rights abuses carried out by the separatists, actively supports terrorism.

While appreciating the urgency of a call to peace, many scientists and academics themselves have reservations about the petition, seeing it as unhelpfully confrontational and even intellectually dishonest. But many have still bravely spoken up for the freedom of expression of the signatories.

Turkey's recently formed Science Academy published a strongly supportive statement in January. "The right to express one's opinions — even if these might be annoying or minority views — is an essential freedom of every citizen and every academic," it said. The academy should know — it was created by those who resigned en masse from the Turkish Academy of Sciences when Erdoğan took it over by decree in 2011. Scientists everywhere should use their pens and send their support. ■

➲ NATURE.COM
To comment online, click on Editorials at: go.nature.com/xhunqv

WORLD VIEW *A personal take on events*

DANIELE FANELLI

# Set up a 'self-retraction' system for honest errors

*Notices should make obvious whether a withdrawal of research is the result of misconduct or a genuine mistake, says* **Daniele Fanelli**.

Self-correction in science has never been so popular and yet so unrewarded. New technologies and a culture of sharing, transparency and public criticism offer an unprecedented opportunity to purge the scientific record of false claims. But retracting those published claims remains a rare and painful process. There are powerful incentives not to do so, for all involved, from universities and scientists to publishers. Retractions still unwittingly punish all who take part. To get the most from self-correction, we must turn blame into praise.

Retractions are a recent tool. The first retraction note recorded in databases was written in 1966 by the authors of a paper on nuclear RNA synthesis. It was an excellent start, but up until ten years ago, retractions were extremely rare, and less than one-fifth of journals had a retraction policy. Today, that proportion has tripled, and retractions are nearing 600 per year.

However, retractions reliably ascribed to honest error account for less than 20% of the total, and are often a source of dispute among authors and a legal headache for journal editors. The recalcitrance of scientists asked to retract work is not surprising. Even when they are honest and proactive, they have much to lose: a paper, their time and perhaps their reputation.

Much reluctance to retract errors would be avoided if we could easily distinguish between 'good' and 'bad' retractions. In our research on misconduct, my colleagues and I informally use terms such as 'honest retraction'. However, these carry a judgement inappropriate for formal notices. Using a more neutral term such as 'withdrawal' could solve that, but it is probably too late to impose a new word on the scientific system.

OUR COMMON **MISSION** IS TO KEEP THE **LITERATURE TRUTHFUL** AND **RELIABLE.**

A more realistic solution is to mimic the way in which bibliometrics researchers use the term self-citation. Superficially, citations all look the same, and are classified as such in databases. However, citations that authors direct at their own work are a self-evident subcategory, which is easily and objectively marked out in any analysis. We can do the same with retractions.

Simply, we should define a self-retraction as any retraction notice that is signed by all co-authors. This is a natural category, which academics, administrators, policymakers and journalists could use unambiguously. Already, retractions resulting from honest error are typically signed by all authors (and most journals require this to avoid legal disputes). Conversely, authors responsible for misconduct add their names to retraction notes only rarely.

To remove ambiguities, journal policies should allow authors to sign only retractions that the researchers have solicited spontaneously, because of a documentable flaw. In all other cases, retraction notes should not be signed — at least not by the authors recognized as responsible for misconduct.

➲ NATURE.COM
Discuss this article online at:
go.nature.com/jrxax2

As long as retraction notes includes in the title a list of all the original authors, as they often already do, their status will be self-evident. If an adjudication of misconduct is disputed in court, as is increasingly the case, then journals could keep the retraction on hold and issue an ordinary expression of concern until the matter is settled.

Self-retractions should be considered legitimate publications that scientists would treat as evidence of integrity. Self-retractions from prestigious journals would be valued more highly, because they imply that a higher sacrifice was paid for the common good. Scientists who committed misconduct would be unable to benefit. Their co-authors — culpable for unwittingly overlooking a fraud — could display their retractions if they wished, but would be unable to claim them as true self-retractions.

Some may argue that such a policy would prompt dishonest researchers to pre-emptively request a retraction, and thereby earn undue praise while sneakily avoiding a future allegation. This is unlikely to be a real problem. Self-retractions would need to be justified by the authors, who would have to provide evidence of the honesty of the mistake. Even if authors fabricated such evidence to conceal a fraud, they could never get away with self-retracting multiple misdeeds. Signing one or two self-retractions may be a badge of honour, but producing many would raise obvious suspicions and mark an author's work as unreliable. Researchers who repeatedly published and self-retracted would be the object not of praise, but of ridicule.

Thus, in the worst-case scenario, it would be only authors who have falsified one or two papers who might benefit from dishonestly self-retracting. Should that be considered a problem? Scientists who remove their flawed work from the literature are sparing the community wasted research and the costs of misconduct investigations. It is in everybody's interest to encourage them to do so, irrespective of their motivations.

Punishment is a means to an end. If praise and reward yield better results, we should enforce them and wish for nothing more. Our common mission is to keep the literature truthful and reliable, and to accomplish that we should be pragmatic, not moralistic. It would not be unholy to grant a year of 'scientific jubilee', during which journal editors allowed authors to self-retract papers, no questions asked. The literature would be purged, repentant scientists would be rewarded, and those who had sinned, blessed with a second chance, would avoid future temptation. ■

**Daniele Fanelli** *is senior research scientist at the Meta-Research Innovation Centre at Stanford University, California.*
*e-mail: email@danielefanelli.com*

# RESEARCH HIGHLIGHTS

*Selections from the scientific literature*

GENETICS

## Modified CRISPR tags RNA in cell

A popular gene-editing technique called CRISPR–Cas9 has been adapted to bind to and track RNA in living cells.

Many labs have adopted CRISPR–Cas9 as a way to edit DNA, using a 'guide RNA' to direct the Cas9 enzyme to the specific DNA sequence to be cut. Gene Yeo of the University of California, San Diego, and his colleagues expanded the technique to target RNA. The team inactivated the Cas9 enzyme that normally slices DNA, and fused it to a fluorescent protein. They then provided a modified guide RNA that directed the disabled enzyme to bind to RNAs.

The approach allowed them to track the fluorescently tagged RNA in living cells, without hindering its movement and function.

***Cell*** **http://doi.org/bdg7 (2016)**

GENOMICS

## Finding our inner Denisovan

Traces of DNA from Denisovans, an extinct group of archaic humans from Asia, have been found in modern humans from Papua New Guinea and elsewhere in Melanesia.

Studies have shown that all non-Africans owe about 2% of their ancestry to Neanderthals, but only Melanesians seem to harbour substantial levels of Denisovan DNA as well. To better characterize the Denisovan heritage of modern humans, Joshua Akey at the University of Washington in Seattle, Svante Pääbo at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, and their team sequenced the genomes of 35 people from Melanesia, and analysed the genomes of another 1,496 people from around the world.

They found that the Melanesians derived between 1.9% and 3.4% of their ancestry from Denisovans. Long stretches of the genomes that were devoid of both Denisovan and Neanderthal DNA included genes that are expressed in certain parts of the brain and one, *FOXP2*, that is involved in speech and language.

***Science*** **http://dx.doi.org/10.1126/science.aad9416 (2016)**

REX LISMAN/GETTY

ANIMAL BEHAVIOUR

## Snakes strike with super speed

Both venomous and non-venomous snakes can strike faster than mammalian prey and predators can react.

David Penning and his colleagues at the University of Louisiana at Lafayette analysed the defensive strikes of non-venomous Texas ratsnakes (*Pantherophis obsoletus*) and two species of venomous snake: western cottonmouth vipers (*Agkistrodon piscivorus leucostoma*; pictured) and western diamond-backed rattlesnakes (*Crotalus atrox*). They found that all snakes could accelerate at more than 160 metres per second squared ($\text{ms}^{-2}$) and reach speeds approaching 3 metres per second. This enables the animals to cover average distances of 13.6–16.7 centimetres in 66–74 milliseconds.

The highest recorded accelerations of nearly 300 $\text{ms}^{-2}$ from a ratsnake and a rattlesnake were roughly 10 times those of jackrabbits attempting to escape.

***Biol. Lett.*** 12, **20160011 (2016)**

PLANETARY SCIENCE

## A peek at Pluto's rich landscapes

Data collected by NASA's New Horizons probe during its Pluto fly-by last year has revealed just how geologically active Pluto is, and that its moon Charon was once active but is now dead.

Jeffrey Moore at the NASA Ames Research Center in Moffett Field, California, and his team report a huge 870,000-$\text{km}^2$ basin on Pluto's surface that contains moving ice layers. It is about 10 million years old at most, and is probably still active. Ancient craters elsewhere on Pluto seem to be up to 4 billion years old and show evidence of tectonics and glacial flow. By contrast, Charon is not active, although it seems to have experienced heavy volcanic activity around 4 billion years ago.

In another paper, William Grundy at the Lowell Observatory in Flagstaff, Arizona, and his team report that methane, carbon monoxide and nitrogen ices are sublimating, condensing and flowing on the surface of Pluto.

***Science*** **http://doi.org/bdg8; http://doi.org/bdg9 (2016)**

MICHAEL SVOBODA/GETTY

CROP SCIENCE

## Plant banks miss crucial seeds

Wild relatives of crop plants are largely missing from seed banks and plant repositories designed to protect biodiversity.

These relatives have genes that could be used to increase yields of crops or make them more resilient. Nora Castañeda-Álvarez at the International Center for Tropical Agriculture in Cali, Colombia, and her colleagues looked at 1,076 relatives of 81 crops, and found that only 45 relatives were adequately represented in seed banks. None of the 81 crops had its wild relatives sufficiently represented in such banks.

The authors call for systematic collecting of wild relatives, and highlight cassava, potato and sorghum as among the highest-priority crops.

***Nature Plants*** **http://dx.doi.org/10.1038/nplants.2016.22 (2016)**

ENGINEERING

## Artificial eye sees in the dark

Taking inspiration from the eyes of an unusual fish, researchers have created a device that can improve the ability of cameras to capture images in low light.

The elephantnose fish (*Gnathonemus petersii*), is known for its low-light vision, and its retina has many reflective microscopic cups that collect light. Hongrui Jiang and his colleagues at the University of Wisconsin–Madison made an array of microscopic cups from glass, lined with reflective aluminium. By transferring the cups to a stretched silicone polymer sheet, the authors shaped the array into a retina-like dome (**pictured**). The cups concentrated incoming light, boosting the sensitivity of the eye by more than three times compared to cameras that did not use this device.

HONGRUI JIANG

The technique blurs the picture slightly, so the team applied an algorithm to sharpen it. The low-cost technique could enhance electronic sensors, which are reaching the limits of their sensitivity, and could have uses in night-vision robots, endoscopes and telescopes, say the authors.

***Proc. Natl Acad. Sci. USA*** **http://doi.org/bdhc (2016)**

CANCER BIOLOGY

## Cancer cells get care packages

Healthy cells that surround a tumour supply it with metabolites that support the voracious appetite of cancer cells — and could one day be targeted by therapeutics.

Cells can swap molecules by producing membrane-bound sacs called exosomes, which act as shuttles between cells. Deepak Nagrath of Rice University in Houston, Texas, and his team studied exosomes from normal connective-tissue cells associated with prostate and pancreatic cancers. They found that the activity of certain metabolic pathways increased in cancer cells when they took up exosomes from nearby normal cells.

The exosomes transported amino acids, lipids and other metabolites to the cancer cells. This cargo helped to sustain tumour growth when nutrients were limiting.

***eLife*** **http://dx.doi.org/10.7554/elife.10250 (2016)**

BIOMECHANICS

## Right prosthetic legs have the edge

An athlete racing with a left-leg prosthesis might run more slowly than a competitor with a right-side prosthesis (**pictured**) at the Paralympic Games, because races are run in the anticlockwise direction.

The speed at which people run round a curve is thought to be limited by the forces exerted by the leg on the inside of the curve. To test this, Paolo Taboga of the University of Colorado Boulder and his team measured the running speeds of 11 athletes wearing leg prostheses. Those with one prosthetic leg were, on average, 3.9% slower when their prosthetic leg was on the inside of the curve, compared with when it was on the outside.

All track events at the Paralympic Games are run on an anticlockwise track, and so athletes with right-leg prostheses may have an advantage over those with left-leg ones.

***J. Exp. Biol.*** 219, **851–858 (2016)**

PLANETARY SCIENCE

## Cassini aids hunt for Planet Nine

Researchers using the Cassini spacecraft have narrowed down the search for the Solar System's hypothetical ninth planet.

Planet Nine is thought to be orbiting in the far outer Solar System, but has not yet been found. If it exists, its gravity should tug slightly on the outer planets, including Saturn. To determine Saturn's orbit, Agnès Fienga at the University of Nice Sophia Antipolis in Valbonne, France, and her colleagues used data from the ground-based radio-dish network that tracks the position of Cassini, which has been orbiting Saturn since 2004. Using a Solar System model refitted with Planet Nine, the researchers conclude that if it exists, the planet probably lies along a particular 21° slice of its predicted orbit.

If Cassini continues to operate until 2020, data from the spacecraft would further improve the estimate of the planet's location, the team says.

***Astron. Astrophys.*** 587, **L8 (2016)**

NEUROSCIENCE

## Forgetting inhibits new memories

Suppressing unwanted memories could interfere with the creation of new ones.

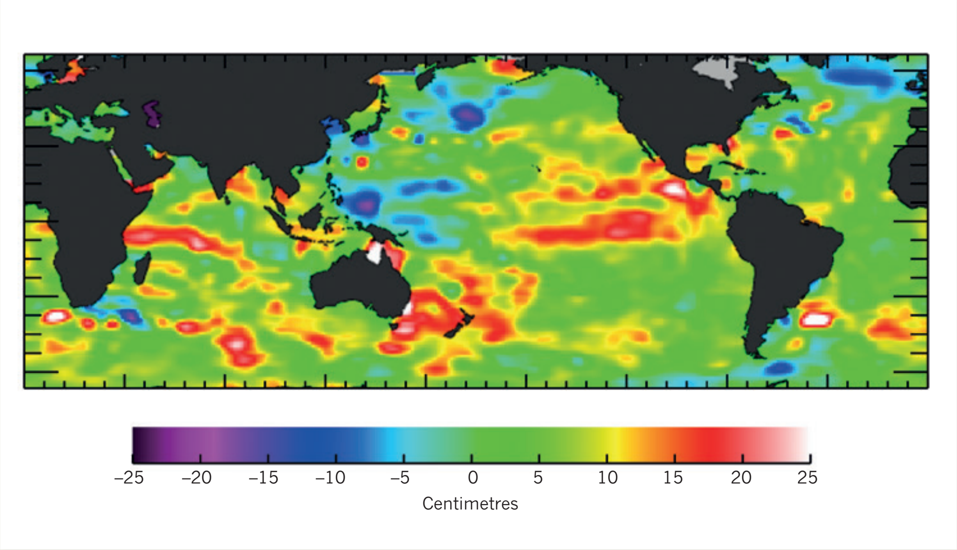
Michael Anderson of the University of Cambridge, UK, and his colleagues asked volunteers to learn pairs of words such as 'pump' and 'oil'. The participants were then cued by one word and asked to either recall, or purposely not think about, the other word in the pair. Between these trials, people viewed various scenes and had to imagine how an object came to be in the scene. The researchers found that, compared with no memory suppression, participants accurately recalled the object's identity about 45% less frequently if the scene was presented shortly before or after memory-suppression trials. The extent of the forgetting effect correlated with how much hippocampus activity was dampened during the memory-suppression trials. (The hippocampus is an area of the brain that is involved in memory processing.)

This 'amnesic shadow' could help to explain memory lapses that can follow traumatic experiences, when people try to suppress certain memories, the authors suggest.

***Nature Commun.*** 7, **11003 (2016)**

➲ NATURE.COM
For the latest research published by *Nature* visit:
www.nature.com/latestresearch

# SEVEN DAYS *The news in brief*



NASA/JPL-CALTECH/OCEAN SURFACE TOPOGRAPHY SCIENCE TEAM

## Satellite tracks sea-surface levels

Jason-3, a joint US–European satellite mission to monitor sea-level rise, has produced its first complete global map of sea-surface height anomalies. The image, released on 16 March, draws on ten days of data collected by the probe after it reached its intended orbit following its 17 January launch. NASA and its European partners — the French space agency CNES and the meteorological agency EUMETSAT — plan a three-year mission for Jason-3. The data that the probe collects could help to improve weather and climate forecasts globally.

FUNDING

### Boost for AI

South Korea announced on 17 March that it would invest 1 trillion won (US$863 million) in artificial-intelligence (AI) research over 5 years. The announcement came two days after Google DeepMind's Go program AlphaGo beat grandmaster Lee Sedol 4–1 in an exhibition match, a feat that prompted a spate of newspaper headlines concerned that South Korea was falling behind in a crucial growth industry. It is not immediately clear whether the cash represents new funding or is money that had been previously allocated to AI efforts. The investment includes a high-profile, public–private research centre near Seoul. See go.nature.com/h3erre for more.

RESEARCH

### Particle promise

Hints of a mysterious particle at the Large Hadron Collider (LHC) near Geneva, Switzerland, just got a little stronger. In December, physicists announced that they had seen an excess of pairs of γ-ray photons — possibly a sign of a particle not predicted by the standard model of physics. The data came from ATLAS and CMS, the two largest detectors at the LHC. A fresh analysis reported on 17 March at a conference in La Thuile, Italy, slightly increases the statistical significance of the signal seen by the CMS experiment. See go.nature.com/pbxwl2 for more.

EVENTS

### Virus control

Mosquitoes carrying a gene that kills their offspring should be released in small pilot studies during the current Zika outbreak, the World Health Organization (WHO) Vector Control Advisory Group said on 18 March. The group also endorsed field tests of mosquitoes carrying bacteria that reduce the insects' ability to transmit Zika, dengue and other viruses. Studies show that these interventions can reduce populations of *Aedes aegypti* mosquitoes, but they have not established whether such interventions can minimize disease burden in humans — a crucial gap that pilot studies could fill, the group said.

### Offshore plans

The US Department of the Interior announced on 15 March that it will not pursue offshore oil and gas development along the US Atlantic coast. The decision, released as part of the offshore-leasing programme for 2017–22, reverses an earlier proposal to sell leases along the central and southern Atlantic coast, from Virginia to Georgia. The proposed plan would move forward with leasing in the Gulf of Mexico and off the Alaskan coast.

### Ebola flare-up ends

Sierra Leone is clear from Ebola after a recent flare-up, the World Health Organization (WHO) announced on 17 March. The date marked 42 days since the last person confirmed to have Ebola virus disease in Sierra Leone tested negative for a second time. The WHO classes this window — twice the incubation period of the virus — as the time needed to declare an outbreak over. On the same day, the WHO confirmed two new cases of Ebola in a rural village in Guinea, which had been declared free of Ebola in December.

### Coral concerns

Faced with ongoing damage to its corals, the Great Barrier Reef Marine Park Authority raised its response status to level three — the highest possible, meaning 'severe regional bleaching' — on 20 March. The decision was made despite heavy rains, which have lowered the high water temperatures that have been killing corals along parts of Australia's northeast coast. According to the authority, divers have found up to 50% coral mortality near the tip of Cape York, but most surveyed sections of the park are in a better state. This regional damage is part of a global coral-bleaching event linked to warmer-than-average waters (see *Nature* http://doi.org/bdmn; 2015).

PEOPLE

## Cancer moonshot

Cancer survivor Gregory Simon was named executive director of the US Cancer Moonshot Task Force on 18 March. Simon was founding president of FasterCures, a non-profit organization in Washington DC dedicated to speeding the development of new medicines, and has worked in government and the pharmaceutical industry. He most recently helmed a health-care investment company. The moonshot project, driven by US vice-president Joe Biden, aims to double the pace of cancer research. President Barack Obama has asked Congress for US$755 million to fund the endeavour in 2017.

## Turkey arrests

Three Turkish academics in Istanbul were taken in for questioning on 14 March and have now been formally arrested and held without bail. They are under suspicion of 'making terrorist propaganda', after signing a petition in January calling for an end to violence between government forces and Kurdish separatists in Turkey's southeast. Those held (**pictured**, from left) are psychologist Esra Mungan from Boğaziçi University, political scientist Muzaffer Kaya from Nişantaşı University and mathematician Kıvanç Ersoy from Mimar Sinan University. Four days earlier, the three had held a press conference affirming their commitment to the 'academics for peace' petition. See go.nature.com/m7uacx for more.

BEYZA KURAL/BIANET

FACILITIES

## China hunts waves

China's plans to detect gravitational waves took a step forward on 20 March, as construction began on the TianQin project at Sun Yat-sen University in Zhuhai. A research building, observation station and ultra-quiet cave laboratory will be built ahead of the launch of three Earth-orbiting satellites. Gravitational waves should be revealed as disturbances in laser beams bounced between the spacecraft. China also has another space-based gravitational-wave detector in the works: the Taiji project from the Chinese Academy of Sciences will involve a trio of Sun-orbiting satellites. See go.nature.com/rral31 for more on China's plans.

## Nobel Prize home

A Nobel Center will be built on the Blasieholmen peninsula in Stockholm's central waterfront by 2019. The City Planning Committee approved detailed plans for the site on 16 March. At a cost of 1.2 billion Swedish kronor (US$146 million), the building will house the Nobel Museum and future prize ceremonies. The centre will be open to visitors and scientists, incorporating research projects, educational efforts, conferences and a library. But its large size and location has attracted criticism, drawing concerns that it will ruin Stockholm's skyline.

## Boaty McBoatface

The British public has flocked to a competition to name a new UK polar research vessel, which is under development and set to be finished in 2019. The National Environment Research Council (NERC) allowed people to submit and vote on names for the £200-million (US$289-million) ship, but its website crashed repeatedly after the poll was launched on 17 March. As *Nature* went to press, the proposed name RRS (Royal Research Ship) *Boaty McBoatface* was the most popular suggestion, with thousands of votes more than the second choice, RRS *Henry Worsley*. The final name will be chosen by NERC.

## COMING UP

**28–29 MARCH**
The US National Institutes of Health holds a special meeting to discuss the spread of Zika virus in the Americas, and ways to contain it.
go.nature.com/vov9qk

**29–31 MARCH**
The University of Central Lancashire in Preston, UK, hosts a meeting of academics and industrialists to discuss applications of nanotechnology.
go.nature.com/ipysuj

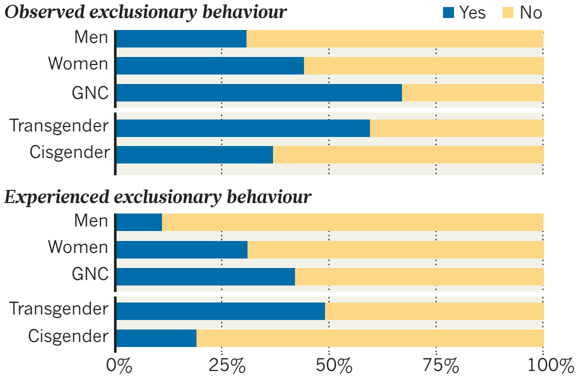
**30 MARCH–22 APRIL**
NORDITA, the Nordic Institute for Theoretical Physics, hosts a meeting in Stockholm on advances in string theory and gauge theory.
go.nature.com/9omyzv

## TREND WATCH

Around one-fifth of physicists from sexual and gender minorities (LGBT) surveyed by the American Physical Society (APS) said that they had been ignored, intimidated or harassed at work (see go.nature.com/4zobly). Transgender respondents were the most affected: almost half had experienced exclusionary or harassing treatment in the previous year. And more than three times as many women as men had experienced such behaviour. See go.nature.com/hitnez for more.

**PHYSICS STRUGGLES WITH INCLUSIVITY**

People from sexual or gender minorities (LGBT) continue to face exclusionary behaviour in the physics community.



GNC, gender non-conforming.

SOURCE: APS

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

# NEWS IN FOCUS

RAY COLLINS/BARCROFT MEDIA/GETTY

**Oceans can be monitored with increasing scope and quality with the use of Argo floats.**

CLIMATE

## Next-generation robotic probes scour the seas

*Initiatives aim to measure global warming's impact on high seas and deep currents.*

**BY JEFF TOLLEFSON**

The Southern Ocean guards its secrets well. Strong winds and punishing waves have kept all except the hardiest sailors at bay. But a new generation of robotic explorers is helping scientists to document the region's influence on the global climate. These devices are leading a technological wave that could soon give researchers unprecedented access to oceans worldwide.

Oceanographers are already using data from the more than 3,900 floats in the international Argo array. These automated probes periodically dive to depths of 2,000 metres, measuring temperature and salinity before resurfacing to transmit their observations to a satellite (see 'Diving deeper'). The US$21-million Southern Ocean Carbon and Climate Observations and Modeling Project (SOCCOM) is going a step further, deploying around 200 advanced probes to monitor several indicators of sea-water chemistry and biological activity in the waters around Antarctica. A primary aim is to track the prodigious amount of carbon dioxide that flows into the Southern Ocean.

"The Southern Ocean is very important, and it's also very poorly known because it's just so incredibly miserable to work down there," says Joellen Russell, an oceanographer at the University of Arizona in Tucson and leader of SOCCOM's modelling team.

Scientists estimate that the oceans have taken up roughly 93% of the extra heat generated by global warming, and around 26% of humanity's $CO_2$ emissions, but it is unclear precisely where in the seas the heat and carbon go. A better understanding ▶

of the processes involved could improve projections of future climate change.

SOCCOM, which launched in 2014, has funding from the US National Science Foundation to operate in the Southern Ocean for six years. Project scientists' ultimate goal is to expand to all the world's oceans. That would require roughly 1,000 floats, and would cost an estimated $25 million per year.

Interest in this global array, dubbed the Biogeochemical Argo, is growing. The Japanese government has put a proposal to expand use of SOCCOM probes on the agenda for the meetings of the Group of 7 leading industrialized nations in Japan in May. And the project is gaining high-level attention as a result: the SOCCOM team has briefed John Holdren, science adviser to US President Barack Obama.

Project scientists are rushing to develop a plan to expand use of the next-generation probes. "It's like, 'Oh, couldn't they wait a year?'" jokes SOCCOM associate director Ken Johnson, an ocean chemist at the Monterey Bay Aquarium Research Institute in Moss Landing, California. His team is drafting a proposal to present to the international Argo steering committee at a meeting that begins on 22 March.

Meanwhile, another set of researchers hopes to extend the existing Argo array beyond its current 2,000-metre limit. The US National Oceanic and Atmospheric Administration (NOAA) is spending about $1 million annually on a Deep Argo project to monitor ocean temperature and salinity down to 6,000 metres. The agency deployed nine Deep Argo floats south of New Zealand in January, and is planning similar pilot arrays in the Indian Ocean and the North Atlantic.

The deep-ocean data will be particularly useful in improving how models simulate ocean circulation, says Alicia Karspeck, an ocean modeller at the National Center for Atmospheric Research in Boulder, Colorado. "From a scientific perspective, it's a no-brainer," she says — noting that the new floats are a low-risk investment compared with spending money on developing models without additional oceanographic data.

NOAA is using two different models of float, both designed to withstand the crushing pressures at the bottom of the sea. And Argo teams in Japan and Europe are already using upgraded floats that can reach down to 4,000 metres. The goal is to establish a new international array of some 1,250 deep-ocean floats — most of which would need to dive to 6,000 metres. Doing so would provide basic data on 99% of the world's seawater.

"We are really still working the bugs out of the equipment and trying to show that we can do this," says Gregory Johnson, a NOAA oceanographer in Seattle, Washington, and one of the principal investigators for Deep Argo.

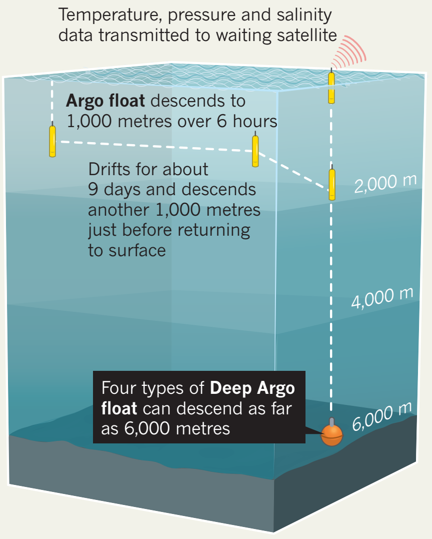
Even if scientists succeed in expanding next-generation ocean probes around the globe, he says, the data that they provide will not supplant detailed measurements of carbon, water chemistry, salinity and temperature that are currently made by ship-based surveys. Deep Argo measures only temperature and salinity, and the technology used in Biogeochemical Argo is not yet sensitive enough to measure subtle changes in the deep ocean.

Still, ship surveys — which are done on average every ten years — cannot follow how heat is taken up by the deep ocean. By contrast, Deep Argo would allow researchers to continually watch heat move through the oceans. That could lead to a better understanding of how the oceans respond to global warming — and how the climate responds to the oceans.

"This has all kinds of ramifications for ecosystems and climate," says Johnson of NOAA. ■

## DIVING DEEPER

Roughly 3,900 Argo floats monitor conditions in the upper oceans. Now, scientists hope to go deeper to cover 99% of Earth's seawater.


Temperature, pressure and salinity data transmitted to waiting satellite
**Argo float** descends to 1,000 metres over 6 hours
Drifts for about 9 days and descends another 1,000 metres just before returning to surface
2,000 m
4,000 m
Four types of **Deep Argo float** can descend as far as 6,000 metres
6,000 m

TECHNOLOGY

# Mobile-phone health apps deliver data bounty

*Smartphone programs allow researchers to recruit large numbers of participants and monitor them in real time.*

**BY ERIKA CHECK HAYDEN**

Last summer, physician Yvonne Chan wondered how the wildfires raging through Washington state were affecting people with asthma — for whom smoke and heat can trigger breathing difficulties. So she tapped into data collected through the Asthma Health iPhone app, which 8,700 people with asthma use each day to record their symptoms and triggers. Chan found that when fires flared up, so did asthma symptoms and reports of environmental triggers among users living near the blazes.

"In the past, stuff like this was just logistically impossible to do," says Chan, director of digital health at the Icahn Institute for Genomics and Multiscale Biology at Mount Sinai in New York City. "It opens up a brand-new area of research."

Smartphone apps developed by academics, pharmaceutical companies and technology giants are making possible large studies that collect real-time data on people's location, environment and health. Last March, for example, Apple debuted its ResearchKit developer tools. Scientists and companies have used these tools to make iPhone apps targeted to specific conditions.

Researchers are only now gaining access to

AMI VITALE/PANOS

these mobile-collected data sets, but the scope of such programs is expanding: on 21 March, Apple announced that ResearchKit apps can now import a user's genetic data from the consumer testing service 23andMe, based in Mountain View, California. Chan's asthma app and MyHeartCounts, a cardiovascular-disease app developed by Stanford University in California, will be the first to be able to easily incorporate users' 23andMe data, if participants allow it.

"The biggest thing we're likely to get with these apps is scale," says neuroscientist James Beck, vice-president for scientific affairs at the Parkinson's Disease Foundation in New York City. mPower, a ResearchKit app that is aimed at people with Parkinson's disease, has enrolled more than 6,800 participants — 3 times the number in the largest previous Parkinson's study.

Collecting health data through an app also makes it easy to share them with other researchers. The mPower and asthma apps ask users whether they would like to make their data available for further analyses. So far, most have agreed — 75% with mPower and 90% in the asthma study. On 3 March, roughly one year after mPower launched, it released a data set from its users — a huge departure from the past, when it would have taken years to collect and distribute such a large amount of data.

"One of the things I like about these new digital ways of gathering information is that it's an opportunity to have data-sharing enabled by the trial itself," says Stephen Friend, president of Sage Bionetworks, a non-profit organization in Seattle, Washington, that developed mPower.

Still, many researchers are taking a wait-and-see approach to the new technology. ResearchKit launched last year with 5 apps; there are now 25, tracking conditions including autism, breast cancer and multiple sclerosis. Some are surprised that more scientists haven't developed their own apps: "I'm shocked to see that we're a year into this and there's so few apps in there," says Atul Butte, director of the Institute for Computational Health Sciences at the University of California, San Francisco.

Almost 9,000 people with asthma use a smartphone app daily to record their symptoms and triggers.

Some may be wary about the quality of the data collected by mobile apps; in many of the ResearchKit studies, study personnel do not meet participants, raising questions about the quality of the data that these participants provide. With this in mind, researchers are working to spot-check their app data. Chan's team has examined whether the height and gender of the asthma-app users correlate as expected with peak flow, a measurement of breathing ability that is usually higher for men and for taller people. So far, that relationship has held up. "We don't have participants randomly entering bad data," Chan says.

*"We're just at the edge of this wave."*

But whether research apps can keep users engaged over the long term is an open question. For instance, mPower uses the iPhone's accelerometer and microphone to measure the steadiness of participants' gait and speech, respectively. But only about 1,000 mPower participants have elected to fill out a survey that assesses cognitive function.

"There's a drop-off in interest as things get more difficult to do," Beck says. "There needs to be some value to the user, so that people don't pick it up and play with it for the first hour and never go back to it."

Friend agrees. His company is working on making the mPower app more user-friendly. "When we look back on it, we'll probably go, 'Wow, this is clunky'," he says of the current app.

And researchers are still working out how best to use the data from such programs. Because participants' data are collected in real time and much more often than usual — for example, daily, as opposed to during a quarterly visit to a physician's surgery — one logical application is in clinical trials of therapeutics. Pharmaceutical company Roche, based in Basel, Switzerland, has developed a Parkinson's app that it is using in a study of a new drug.

Others suggest that mobile-enabled research may eventually lead to wearable devices that automatically collect information about participants in real time — such as those being developed by Verily in Mountain View. Its Baseline Study will use wearable devices to collect user data, with the aim of gaining insights into how to detect and prevent disease.

Says Beck: "We're just at the edge of this wave." ■

MARIO TAMA/GETTY

## MORE ONLINE

**TOP STORY**

The Zika virus and birth defects: what we know and what we don't go.nature.com/jylflc

**MORE NEWS**

- GM zebrafish forms technicolour 'skinbow' go.nature.com/7toog9
- Australian cryptologists concerned by restrictive exports law go.nature.com/f3qtcu
- Dengue vaccine aces trailblazing trial go.nature.com/ai6o7c

**NATURE PODCAST**

Toggling brains with radio waves; building stuff that lasts; and thrill-seeking rodents nature.com/nature/podcast

YAN-HONG NIE

A bounty of research monkeys in China is enabling neuroscience to flourish.

CHINA

# Science wins in five-year plan

*Oceanography, brain science and stem cells are among the Chinese research fields that look set to grow by 2020.*

BY DAVID CYRANOSKI

From a slowing economy to geopolitical tensions in the South China Sea, it is a testing time for China's ruling Communist Party. But according to its 13th Five-Year Plan, approved on 16 March, its science aspirations seem to be unbridled.

China already intended its research expenditure to rise to 2.5% of gross domestic product by 2020, from less than 2.2% over the past 5 years.

A draft version of the latest Five Year Plan, as well as statements by key politicians, bolsters the idea that innovation through science and technology is a priority. For some of the themes that are set to shape Chinese research over the next five years, *Nature* spoke to a range of scientists.

### THE OCEAN DEEP

In 2012, 'oceanauts' aboard the research submersible *Jiaolong* descended more than 7,000 metres beneath the waves, marking China's entry into an elite club of nations capable of reaching the hadal zone — the deepest part of the ocean, which begins at 6,000 metres below sea level. Over the next five years, Chinese scientists will build one crewed and one uncrewed submersible, according to a plan released by the science ministry in February, each of which will be able to reach depths of 11,000 metres — the very bottom of the hadal zone.

*"For deep-sea technology, this five years will be a golden period."*

"For deep-sea technology, this five years will be a golden period," says Cui Weicheng of the Hadal Science and Technology Research Center at Shanghai Ocean University.

The uncrewed vessel will be similar to Nereus, the advanced US submersible that imploded in 2014 and will not be replaced. The crewed vessel will hold at least two people, more than the *DEEPSEA CHALLENGER*, which took film director James Cameron on a solo dive to the deepest point of the Mariana Trench in 2012. The hadal zone is one of the most poorly studied habitats on Earth, and is home to mysterious tube worms, sea cucumbers and jellyfish. Researchers are also interested in its role in the carbon cycle, because the microbes there digest a surprising amount of organic matter. Chinese scientists hope to use both submersibles to explore the zone in more detail than before.

Independently of the latest five-year plan, Cui has also developed a 'movable laboratory' composed of three landers, a robotic submersible and a crewed vehicle (W. Cui *et al. Meth. Oceanogr.* **10,** 178–193; 2014).The robotic submersible and first lander were tested down to 4,000 metres last October. A mother ship that controls the robot and landers is due to be launched on 24 March, and the first scientific expedition is planned for August, in the New Britain Trench off Papua New Guinea. Together, these projects "could help shorten the gap" between Chinese ocean science and technology and the most advanced capabilities elsewhere, says Cui.

### BRAIN SCIENCE

The United States, Europe and Japan have each announced their own massive projects to map the brain, and China has had one in the works for several years. The latest five-year plan calls for brain science to be a priority — and most of the resources are expected to be channelled through the China project, which is due to be officially announced shortly, say Chinese researchers.

The brain project is expected to focus on brain disease, in particular through studying animal models, and on artificial intelligence. Scientists in China acknowledge that they are far behind the rest of the world in terms of top-level talent in brain science, but several factors could enable them to catch up. China's neuroscience community is growing — the Chinese Neuroscience Society now has 6,000 members, compared to just 1,500 ten years ago — and the country has hundreds of thousands of research monkeys. Furthermore, China's tens of millions of patients with psychiatric or degenerative brain disease will facilitate clinical studies.'

The research monkeys have already allowed Chinese researchers to take the lead in using gene-editing technologies to produce models of autism spectrum disorder and other conditions. The bounty of research animals is also starting to draw interest from abroad — a primate research centre in Shenzhen is being jointly established with the Cambridge-based Massachusetts Institute of Technology.

### CONSERVATION CORRIDORS

With actor Jackie Chan and basketball star Yao Ming involved in campaigns attacking the trade in protected animals such as bears, which are milked for their bile, and elephants, targeted for their ivory, conservation has become a high-profile issue in China.

The latest five-year plan will launch efforts to protect the giant panda, tiger and Asian elephant in the wild, says Zhang Li, a conservation biologist at Beijing Normal University. "There will be a big budget to restore habitat for these species," says Zhang. The projects will focus on corridors between protected areas that greatly increase habitats by letting the animals move from one reserve to another.

A biodiversity hotspot between Laos, Myanmar and the southwestern Chinese province of Yunnan requires protection in particular, says Stuart Pimm, a biodiversity specialist at Duke University in Durham, North Carolina. The forest there has been converted into rubber plantations, he says, "and the level of hunting is worse than any place I've ever been". But a focus on protecting pandas, elephants and tigers could leave other animals at risk, he pointed out in November (B. V. Li and S. L. Pimm *Conserv. Biol.* **30,** 329–339; 2016).

### STEM CELLS

In the wake of the five-year plan, China will gain a funding initiative called 'Stem Cell and Translational Research', according to stem-cell researchers Pei Gang, president of Tongji University in Shanghai, and Pei Duanqing, director of the Guangzhou Institutes of Biomedicine and Health. The stem-cell programme will be one of the first to award grants under a new competitive review and evaluation process, replacing a system that critics said rewarded scientific and political connections rather than merit. Following the previous five year plan, China invested roughly 3 billion yuan (about US$460 million) in stem-cell research.

The pair says that there will be a big increase over the next five years but did not give exact figures. "Given the size of its population and the wide spectrum of unmet medical needs, China recognizes the promise of stem-cell and regenerative medicine as one of the key thrusts for modernizing its medical-service system," says Pei Gang.

OLLI GEIBEL/REX/SHUTTERSTOCK

Smog hits Beijing, a pollution black spot, on 25 December 2015 — for the fourth time that month.

### POLLUTION CONTROL

In a country that places great value on social harmony, air and water pollution are the trigger for an increasing number of protests.

Under a plan that began in 2012, the government is already trying to reduce the levels of airborne particulate matter measuring less than 2.5 micrometres across ($PM_{2.5}$), which is small enough to penetrate deep into the respiratory system. By 2017, it wants to achieve reductions of 25% in the Beijing area, 20% in the Yangtze River Delta and greater Shanghai area, and 15% in the Pearl River Delta. Major nationwide environmental initiatives outlined in the latest five-year plan will tackle transportation, clean energy and environmental protection, says Wei-xian Zhang, director of the State Key Lab for Pollution Control at Tongji University.

The government will also target pollution black spots, such as smog in Beijing and fertilizer pollution in Lake Tai near Shanghai. Funding to control air pollution alone will increase by at least four times, says Zhang, and several new national laboratories focusing on clean energy and environmental research have also been funded for the next five years. "China is and will continue to be the largest market in air-, soil- and water-pollution control technologies," says Zhang. "To some degree, the whole country will be a huge laboratory for environmental research, such as smog mitigation." ■

CLIMATE

# China's carbon emissions could peak sooner than forecast

*Five-year plan advances policy to reduce reliance on coal and expand renewable energy.*

BY JEFF TOLLEFSON

The world's largest greenhouse-gas emitter is turning a corner on climate change. China's 13th Five-Year Plan reinforces the country's seismic shift away from dirty coal, and many specialists now think that Chinese emissions are already nearing their peak — well ahead of schedule.

Approved on 16 March, the plan sets out basic goals and requirements for energy use and the environment until 2020 — and establishes an overarching strategy for economic development, as well as some themes to shape the direction of research (see opposite). In particular, the document strengthens mandatory targets put in place over the past decade to reduce energy use, curb air pollution and promote the development of wind, solar and nuclear power.

These efforts have begun to work: China's coal consumption declined by an estimated 3.7% in 2015, according to statistics released in February by the Chinese government.

Such a decrease is unprecedented, says Barbara Finamore, Asia director for the Natural Resources Defense Council, an environmental-advocacy group headquartered in New York City. "I think it's catching everyone by surprise."

The new plan calls for an 18% reduction in carbon intensity, which is a measure of how much carbon dioxide is emitted per unit of gross domestic product. That is slightly stronger than the 17% target set in 2011. The latest plan also seeks to limit the country's total energy use. China consumed energy equivalent to 4.3 billion tonnes of coal in 2015, and the plan would seek to cap that figure at the equivalent of 5 billion tonnes by 2020.

Nonetheless, the document does not specify how China will hit its targets. "The point of this is to set the tone and direction," says Ranping Song, who handles climate ▶

▸ policy in developing countries for the World Resources Institute, an environmental think tank in Washington DC. Song expects China to release detailed plans in coming months about how various sectors of its economy will meet the new commitments.

But China is already on track to achieve — and probably exceed — its previous targets. The latest data suggest that the country may have already halted its dramatic rise in coal use, beating the 2020 deadline that it set 2 years ago. China also leads the world in the deployment of renewable energy, investing some US$110 billion in 2015.

At the United Nations climate summit in Paris last year, China committed to halting growth in greenhouse-gas emissions by 2030, but consensus is building that a peak could come by 2025 — if not sooner. In addition to energy trends, the latest forecasts account for slower economic growth, as well as a shift away from heavy manufacturing and the production of steel and other commodities.

Some fear that coal consumption could spike again, along with carbon emissions, if China's slowing economy revives. But a London School of Economics study published on 16 March concludes that this is unlikely (F. Green and N. Stern *Clim. Pol.* http://doi.org/bdmm; 2016). The Chinese government's latest energy data suggest that emissions may have dropped in 2015, says Fergus Green, a policy analyst who co-authored the study with economist Nicholas Stern. This means that China's emissions may have already peaked.

One big question is whether China can rein in oil use in the growing transportation sector, in which the government has been less aggressive. Nonetheless, Finamore says, strict new requirements on air pollution, driven by rising anger among Chinese citizens, are pushing China in the right direction. "This is the new normal." ■

MIKE BLAKE/REUTERS

A killer whale at SeaWorld, which has stopped breeding the animals in captivity.

MAMMALOGY

# Clash over killer-whale captivity

*Lifespan of animals kept in parks is at centre of dispute.*

BY EWEN CALLAWAY

In a decision hailed by animal-rights groups, the US marine-park company SeaWorld Entertainment announced last week that it will no longer breed killer whales. But whether captivity harms the planet's biggest predator is an area of active scientific debate.

The latest arguments centre on two 2015 studies that drew dramatically different conclusions about the lifespans of captive killer whales (*Orcinus orca*), relative to those of wild populations. Although many factors affect well-being, an apparent discrepancy between the survival of captive and wild animals has long been cited by activists as evidence of the poor welfare of captive killer whales.

One of the studies[1] is authored by a team largely made up of researchers at SeaWorld, which is headquartered in Orlando, Florida, and owns several animal parks that keep killer whales; the other[2] is by two former killer-whale trainers at the company who feature in the 2013 documentary film *Blackfish*, which is critical of SeaWorld. In letters published last week[3,4], authors from each paper accuse the others of cherry-picking data to support positions on whether the animals should be captive — charges that each team in turn rejects.

Although SeaWorld's captive-killer-whale programme now has an expiration date, the company's existing 23 animals will remain in parks for the rest of their lives, and its pregnant female Takara will give birth in captivity. Another 33 animals are held in other marine parks around the world.

Robust studies of killer whales' longevity are needed to improve the well-being of the remaining captive animals, says Douglas DeMaster, science director at the US National Oceanic and Atmospheric Administration's Alaska Fisheries Science Center in Seattle, Washington.

But the annals of research on captive killer whales are slim. Before 2015, the last major published study[5] dates to 1995, when US government scientists calculated that the annual survival rate of captive killer whales was several per cent lower than that of a wild population off the coast of Washington state called southern resident killer whales.

In one of the 2015 studies[2], the former trainers — John Jett, a biologist at Stetson University in DeLand, Florida, and Jeffrey Ventre, a veterinary surgeon at Lakeview Campus Medical Facility in Yakima, Washington — attempted to measure how captive whales have fared since conditions were improved in the 1980s. They pooled data from between 1961 and 2013 on 201 captive killer whales in institutions around the world, including SeaWorld. They concluded that survival rates in captivity have improved since 1985, but that even the most recent survival rates are below those of animals in the wild.

In the other 2015 study[1], researchers led by SeaWorld veterinary surgeon Todd Robeck came to a very different conclusion: that animals now in captivity at SeaWorld's US parks live just as long as wild populations. The researchers looked only at animals held at those parks after 2000, and produced a survival rate that is higher than a rate that they calculated for southern resident killer whales — and equivalent to that of another wild population that lives in the waters off British Columbia, Canada.

Now, each lead author has taken aim at the work of the other. In a letter published in *Marine Mammal Science*[3], Robeck and three colleagues note that Jett and Ventre included in their 2015 study stranded animals, which might have arrived in captivity in poor health, and newborns, which are at particularly high risk of death. This pushes down the apparent survival rate of captive animals, say the researchers.

In the same journal, Jett responds[4] to that critique, and accuses Robeck's 2015 study of bias because, for instance, it compares captive whales to the southern resident population, which is endangered and exposed to pollutants and shipping traffic, and whose numbers have waxed and waned over the past four decades.

Jett says that his and Ventre's study was intended to take a wide look at captive-killer-whale survival, so they included as many data as possible. But Robeck stands by his critique. "They can include all the animals they want," he says. "The conclusions they made were not based on the evidence they showed."

DeMaster notes that the comparison that Robeck and his colleagues made between captive killer whales and a disturbed wild population is not useful. He adds that it is also difficult to compare the approaches taken by the two teams, because they analyse different animals over different periods.

On 8 March, a further group of researchers entered the fray, criticizing the 2015 Robeck study on another front. In the *Journal of Mammalogy*[6], the group charges that Robeck's study implied that evidence for a long post-reproductive lifespan in killer whales is an artefact stemming from overestimated ages of adults in the early days of research on captive killer whales. "People started looking at killer whales in the early 1970s and they weren't immediately experts," says Robeck, who has also published a response[7] to that critique.

***"People started looking at killer whales in the early 1970s and they weren't immediately experts."***

The authors of the critique say that the evidence for the post-reproductive lifespan, a rare evolutionary adaptation otherwise seen only in humans and in pilot whales, is robust. "There are whales still alive now that were around in the 70s that haven't had a calf," says one of the authors, Darren Croft, a behavioural ecologist at the University of Exeter, UK. It will take more observation time to put firm numbers on the post-reproductive lifespan of killer whales, says Andrew Foote, an evolutionary ecologist at the University of Bern and another of the co-authors.

The only way to resolve the dispute over the longevity of captive killer whales is for different teams to analyse the same data in the same manner, says DeMaster. Such studies could improve the well-being of captive animals by, for instance, identifying the facilities and husbandry practices that most benefit them. ■

1. Robeck, T. R., Willis, K., Scarpuzzi, M. R. & O'Brien, J. K. *J. Mammal.* http://dx.doi.org/10.1093/jmammal/gyv113 (2015).
2. Jett, J. & Ventre, J. *Mar. Mamm. Sci.* **31,** 1362–1377 (2015).
3. Robeck, T., Jaakkola, K., Stafford, G. & Willis, K. *Mar. Mamm. Sci.* http://dx.doi.org/10.1111/mms.12278 (2016).
4. Jett, J. *Mar. Mamm. Sci.* http://dx.doi.org/10.1111/mms.12313 (2016).
5. Small, R. J. & DeMaster, D. P. *Mar. Mamm. Sci.* **11,** 209–226 (1995).
6. Franks, D. W. *et al. J. Mammal.* http://dx.doi.org/10.1093/jmammal/gyw021 (2016).
7. Robeck, T. R., Willis, K., Scarpuzzi, M. R. & O'Brien, J. K. *J. Mammal.* http://dx.doi.org/10.1093/jmammal/gyw023 (2016).

**CORRECTION**

The News Feature 'The red-hot debate about transmissible Alzheimer's' (*Nature* **531,** 294–297; 2016) erroneously stated that growth hormone had been derived from the adrenal glands of cadavers. In fact, it came from the pituitary glands.

BY DAVIDE CASTELVECCHI

# THE NEXT WAVE

ILLUSTRATION BY MARK GARLICK

A momentous signal from space has confirmed decades of theorizing on black holes — and launched a new era of gravitational-wave astronomy.

The event was catastrophic on a cosmic scale — a merger of black holes that violently shook the surrounding fabric of space and time, and sent a blast of space-time vibrations known as gravitational waves rippling across the Universe at the speed of light.

But it was the kind of calamity that physicists on Earth had been waiting for. On 14 September, when those ripples swept across the freshly upgraded Laser Interferometer Gravitational-Wave Observatory (Advanced LIGO), they showed up as spikes in the readings from its two L-shaped detectors in Louisiana and Washington state. For the first time ever, scientists had recorded a gravitational-wave signal.

"There it was!" says LIGO team member Daniel Holz, an astrophysicist at the University of Chicago in Illinois. "And it was so strong, and so beautiful, in both detectors." Although the shape of the signal looked familiar from the theory, Holz says, "it's completely different when you see something in the data. It's this transcendent moment".

The signal, formally designated GW150914 after the date of its occurrence and informally known to its discoverers as 'the Event', has justly been hailed as a milestone in physics. It has provided a wealth of evidence for Albert Einstein's century-old general theory of relativity, which holds that mass and energy can warp space-time, and that gravity is the result of such warping. Stuart Shapiro, a specialist in computer simulations of relativity at the University of Illinois at Urbana–Champaign, calls it "the most significant confirmation of the general theory of relativity since its inception".

But the Event also marks the start of a long-promised era of gravitational-wave astronomy. Detailed analysis of the signal has already yielded insights into the nature of the black holes that merged, and how they formed. With more events such as these — the LIGO team is

Binary black holes radiate a huge amount of orbital energy as gravitational waves.

analysing several other candidate events captured during the detectors' four-month run, which ended in January — researchers will be able to classify and understand the origins of black holes, just as they are doing with stars.

Still more events should appear starting in September, when Advanced LIGO is scheduled to begin joint observations with its European counterpart, the Franco–Italian-led Advanced Virgo facility near Pisa, Italy. (The two collaborations already pool data and publish papers together.) This detector will not only contribute crucial details to events, but could also help astronomers to make cosmological-distance measurements more accurately than before.

"It's going to be a really good ride for the next few years," says Bruce Allen, managing director of the Max Planck Institute for Gravitational Physics in Hanover, Germany.

"The more black holes they see whacking into each other, the more fun it will be," says Roger Penrose, a theoretical physicist and mathematician at the University of Oxford, UK, whose work in the 1960s helped to lay the foundation for the theory of the objects. "Suddenly, we have a new way of looking at the Universe."

## A MATTER OF ENERGY

Physicists have known for decades that every pair of orbiting bodies is a source of gravitational waves. With each revolution, according to Einstein's equations, the waves will carry away a tiny fraction of their orbital energy. This will cause the objects to move a bit closer together and orbit a little faster. For familiar pairs, such as the Moon and Earth, such energy loss is imperceptible even on timescales of billions of years.

But dense objects in very close orbits can lose energy much more quickly. In 1974 , radio astronomers Russell Hulse and Joseph Taylor, then of the University of Massachusetts Amherst, found just such a system: a pair of dense neutron stars in orbit around each other. As the years went by, the scientists found that this 'binary pulsar' was losing energy and spiralling inwards exactly as predicted by Einstein's theory.

The two black holes detected by LIGO had probably been losing energy in this way for millions, if not billions, of years before they reached the end. But LIGO did not register the gravitational waves coming from them until 9:50:45 Coordinated Universal Time on 14 September, when the wave's frequency rose above some 30 cycles per second (hertz) — corresponding to 15 full black-hole orbits per second — and was finally high enough for the detectors to distinguish it from background noise.

But then, in just 0.2 seconds, LIGO watched the signal surge to 250 hertz and suddenly disappear, as the black holes made their final 5 orbits, reached orbital velocities of half the speed of light and coalesced into a single massive object (see 'What made the wave').

The LIGO and Virgo teams soon went to work extracting every bit of information possible. At the most fundamental level, the signal gave them an existence proof: the fact that the objects came so close to each other before merging meant that they had to be black holes, because ordinary stars would need to be much bigger. "It is, I think, the clearest indication that black holes are really there," says Penrose.

The signal also provided researchers with the first empirical test of general relativity beyond regions — including the space around the binary pulsar — where there is comparatively little space-time warping. There was no empirical evidence that the theory would keep its validity at the extreme energies of merging black holes, says Shapiro — but it did.

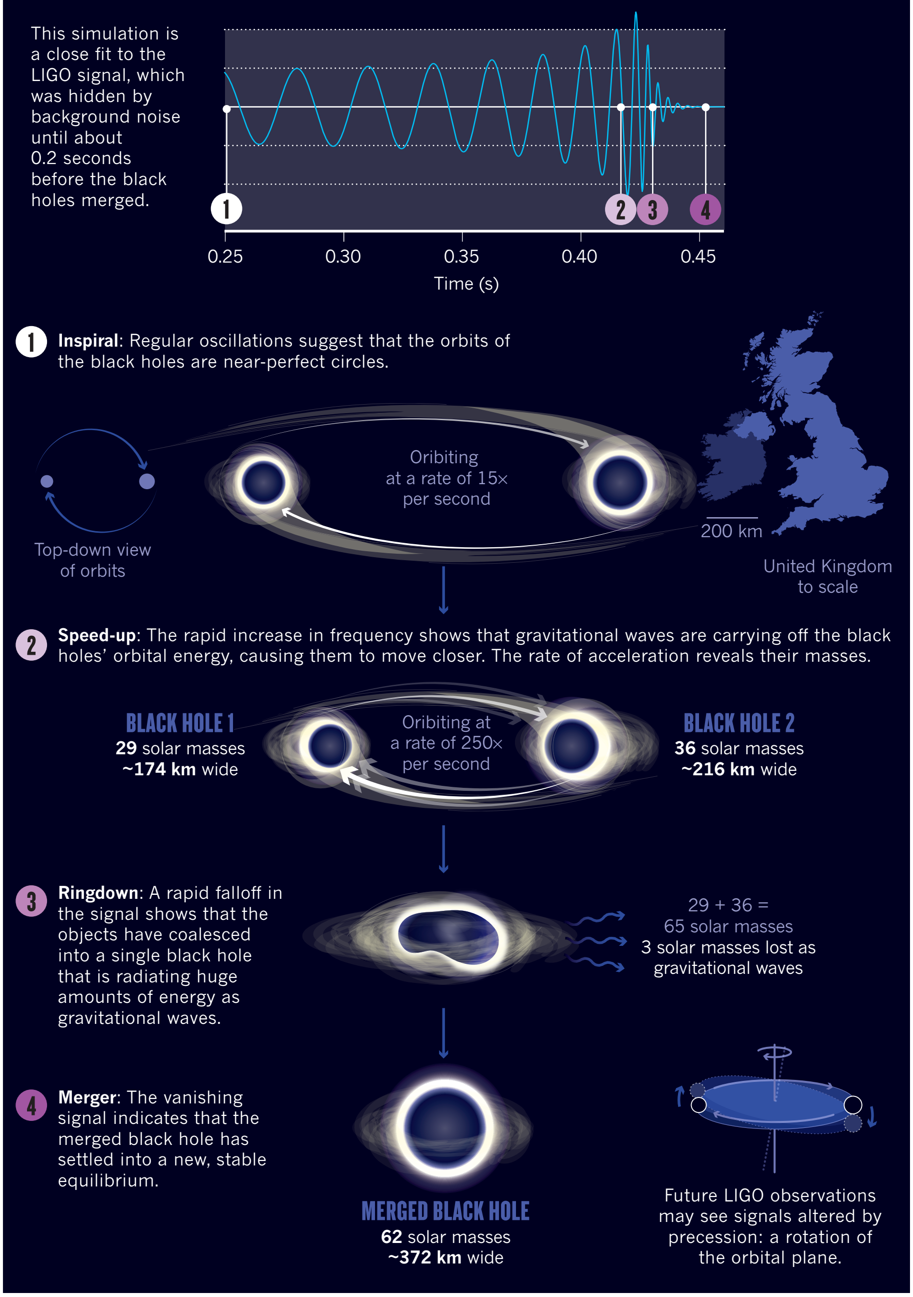
The signal held a trove of more-detailed information as well. By scrutinizing its shape just before the final cataclysm, the scientists found that it closely approximated a simple sine wave with a steadily increasing frequency and amplitude. According to B. S. Sathyaprakash, a theoretical physicist at Cardiff University, UK, and a senior LIGO researcher, this pattern suggests that the orbits of the black holes were nearly circular, and that LIGO probably had a bird's-eye view of the circles, looking almost straight down on them rather than edge-on.

In addition, the LIGO and Virgo teams were able to use the frequency of the observed wave, along with its rate of acceleration, to estimate the masses of the two black holes: because heavier objects radiate energy in the form of gravitational waves at a faster rate than do lighter objects, their pitch rises more quickly.

By recreating the Event with computer simulations, the scientists calculated that the two black holes weighed about 36 times and 29 times the mass of the Sun, respectively, and that the combined black hole weighed about 62 solar masses[1]. The lost difference, about three Suns' worth, was dispersed as gravitational radiation — much of it during what physicists call

## WHAT MADE THE WAVE

The first signal ever detected by the Advanced Laser Interferometer Gravitational-Wave Observatory (LIGO) lasted just 0.2 seconds, but conveyed a wealth of information.



This simulation is a close fit to the LIGO signal, which was hidden by background noise until about 0.2 seconds before the black holes merged.

1 **Inspiral**: Regular oscillations suggest that the orbits of the black holes are near-perfect circles.

2 **Speed-up**: The rapid increase in frequency shows that gravitational waves are carrying off the black holes' orbital energy, causing them to move closer. The rate of acceleration reveals their masses.

3 **Ringdown**: A rapid falloff in the signal shows that the objects have coalesced into a single black hole that is radiating huge amounts of energy as gravitational waves.

4 **Merger**: The vanishing signal indicates that the merged black hole has settled into a new, stable equilibrium.

Future LIGO observations may see signals altered by precession: a rotation of the orbital plane.

SOURCE: REF. 1/NIK SPENCER/NATURE

the 'ringdown' phase, when the merged black hole was settling into a spherical shape. (For comparison, the most powerful thermonuclear bomb ever detonated converted only about 2 kilograms of matter into energy — roughly $10^{30}$ times less.) The teams also suspect that the final black hole was spinning at perhaps 100 revolutions per second, although the margin of error on that estimate is large.

The inferred masses of the two black holes are also revealing. Each object was presumably the remnant of a very massive star, with the larger star approaching 100 times the mass of the Sun and the smaller one a little less. Thermonuclear reactions are known to convert hydrogen in the cores of such stars into helium much faster than in lighter stars, which leads them to collapse under their own weight only a few million years after they are born. The energy released by this collapse causes an explosion called a type II supernova, which leaves behind a residual core that turns into a neutron star or, if it's massive enough, a black hole.

Scientists say that type II supernovae should not produce black holes much bigger than about 30 solar masses — and both black holes were at the high end of that range. This could mean that the system formed from interstellar gas clouds that were richer in hydrogen and helium than the ones typically found in our Galaxy, and that were poorer in heavy elements — which astronomers call metals.

Astrophysicists have calculated that stars formed from such low-metallicity clouds should have an easier time forming massive black holes when they explode, explains Gijs Nelemans, an astronomer at Radboud University Nijmegen in the Netherlands and a member of the Advanced Virgo collaboration. That's because during a supernova explosion, smaller atoms are less likely to be blown away by the blast. Low-metallicity stars thus "lose less mass, so more of it goes into the black hole, for the same initial mass", Nelemans says.

### TWO BY TWO

But how did these two black holes end up in a binary system? In a paper[2] published at the same time as the one reporting the discovery, the LIGO and Virgo teams described two commonly accepted scenarios.

The simplest one is that two massive stars were born as a binary-star system, forming from the same interstellar gas cloud like a double-yolked egg, and orbiting each other ever since. (Such binary stars are common in our Galaxy; singletons such as the Sun are the exception, rather than the rule.) After a few million years, one of the stars would have burned out and gone supernova, soon to be followed by the other. The result would be a binary black hole.

The second scenario is that the stars formed independently, but still inside the same dense stellar cluster — perhaps one similar to the globular clusters that orbit the Milky Way. In such a cluster, massive stars would sink towards the centre and, through complex interactions with lighter stars, form binary systems, possibly long after their transformation into black holes.

Simulations made by Simon Portegies Zwart, an astrophysicist at Leiden University in the Netherlands, show[3] that massive stars are more likely to form in dense clusters, where collisions and mergers are more common. He also finds that once a binary black-hole system forms, the complex dynamics of the cluster's centre would probably kick the pair out at high speed. The binary that Advanced LIGO detected may have wandered away from any galaxy for billions of years before merging, he says.

Although the LIGO and Virgo teams were able to learn a lot from the Event, there is much more that gravitational waves could teach them, even in the case of black-hole mergers. The detectors showed that immediately after the black holes merged, the waves quickly died down as the resulting black hole settled into a symmetrical shape. This is consistent with predictions made by theoretical physicist C. V. Vishveshwara in the early 1970s, a time when "gravitational waves and black holes both belonged to the realm of mythology", he says. "At that time, I had not imagined that it would ever be verified," says Vishveshwara, who is director emeritus of the Jawaharlal Nehru Planetarium in Bangalore, India.

But LIGO saw only just over one cycle of the Event's ringdown waves before the signal became buried once more in the background noise — not yet enough data to provide a rigorous test of Vishveshwara's predictions.

More-stringent tests will be possible if and when LIGO detects black-hole mergers that are larger than this one, or that occur closer to Earth than the Event's estimated distance of 1.3 billion light years, and thus give 'louder' waves that stay above the noise for longer.

Alessandra Buonanno, a LIGO theorist and director of the Max Planck Institute for Gravitational Physics in Potsdam-Golm, Germany, says that a more detailed picture of the ringdown stage could reveal how fast the final black hole rotates, as well as whether its formation gave it a 'natal kick', imparting a high velocity.

In addition, says Sathyaprakash, "we are especially waiting for systems that are much lighter, so they last longer". Such events could include the mergers of lighter binary black holes, of binary neutron stars or of a black hole with a neutron star. Each type would deliver its own signature chirp, and could produce a signal that stays above LIGO's threshold of sensitivity for several minutes or more.

"GW150914 is in some sense a very vanilla system," says Chad Hanna, a LIGO member at Pennsylvania State University in University Park. "It's beautiful, of course, but it doesn't have all the crazy things that one might expect."

## SPACE ARTISTRY

One phenomenon that Sathyaprakash is eager to observe is a 'precession' of the black holes' orbital plane, meaning that their paths trace a kind of 3D rosette. This is a relativistic effect that has no counterpart in Newtonian gravity, and it should produce a characteristic fluctuation in the strength of the gravitational waves. But orbital precession occurs only when two black holes have axes of rotation that point in random directions, and it disappears when the axes are both perpendicular to the orbital plane. The occurrence of a precession could provide clues to how the black holes formed.

It's hard to be sure about that possibility because there are many uncertainties in simulating supernovas. But astrophysicists suspect that parallel spins generally signify that the original two stars were born together out of the same whirling gas cloud. Similarly, they think that random spins result from black holes that formed separately and later fell into orbit around each other. Once the observatories find more mergers, they may be able to determine which type of system occurs more frequently.

Although detecting more events will help LIGO to do lots of science, its interferometers have intrinsic limitations that make it necessary to work together with a worldwide network of similar detectors that are now coming online.

First, LIGO's two interferometers are not enough for scientists to determine precisely where the waves came from. The researchers can get some information by comparing the signal's time of arrival at each detector: the difference enables them to calculate the wave's direction relative to an imaginary line drawn between the two. But in the case of the Event, which recorded a difference of 6.9 milliseconds, their calculations limited the field of possibilities merely to a wide strip of southern sky.

> "IT IS, I THINK, THE CLEAREST INDICATION THAT BLACK HOLES ARE REALLY THERE."

Had Virgo been online, the scientists could have narrowed down the direction substantially by comparing the waves' arrival times at three places. With a fourth interferometer (Japan is building an underground one called KAGRA, for Kamioka Gravitational-Wave Detector, and India has its own LIGO in planning), their precision would improve much more.

Knowing an event's direction will in turn remove one of the biggest uncertainties in determining its distance from Earth. Waves that approach from a direction exactly perpendicular to the detector — either from above or from below, through Earth — will be recorded at their actual amplitude, explains Fulvio Ricci, a physicist at the University of Rome La Sapienza and the spokesperson for Virgo. Waves that come from elsewhere in the sky, however, will hit the detector at an angle and produce a somewhat smaller signal, according to a known formula. There are even some blind spots, where a source cannot be seen by a given detector at all.

Determining the direction will therefore reveal the exact amplitude of the waves. By comparing that figure with the waves' amplitude at the source, which the researchers can derive from the shape of the signal, and by knowing how the amplitude decreases with distance, which they get from Einstein's theory, they can then calculate the distance of the source to a much higher precision.

This situation is almost unprecedented: conventionally, astronomical distances need to be estimated by looking at the brightness of known objects in locations that range from the Solar System to distant galaxies. But the measured brightness of those 'standard candles' can be dimmed by stuff in between. Gravitational waves have no such limitation.

## RAISING THE ALARM

There is another important reason why scientists are eager to have precise estimates of the waves' provenance. The LIGO and Virgo teams have arranged to give near-real-time alerts of intriguing events to more than 70 teams of conventional astronomers, who will use their optical, radio and space-based telescopes to see whether those events produced any form of electromagnetic radiation. In return, the LIGO and Virgo collaborations will be sifting through data to search for gravitational waves that could have been generated by events, such as supernova explosions, seen by the conventional observatories.

Some 20 teams tried to follow up on the Event, mostly to no avail. NASA's Fermi Gamma-ray Space Telescope did see a possible burst of γ-rays about 0.4 seconds later, coming from an equally vague but compatible region of the southern sky[4]. But most observers now consider it to be a coincidence. Such γ-rays could, in principle, have been produced when gas orbiting the binary black hole was heated up during the merger, says Vicky Kalogera, a LIGO astrophysicist at Northwestern University in Evanston, Illinois. But "our astrophysical expectation has been that the gas from stars that formed the binary black hole has long dispersed. There shouldn't be any significant gas around", she says.

Going forward, however, matching gravitational waves with electromagnetic ones could usher in a new era of astronomy. In particular, mergers of neutron stars are expected to produce short γ-ray bursts. Researchers could then measure how far the light from those bursts is shifted towards the red end of the spectrum, which would tell astronomers how fast the stars' host galaxies are receding owing to the expansion of the Universe.

Matching those redshifts to distance measurements calculated from gravitational waves should give estimates of the current rate of cosmic expansion, known as the Hubble constant, that are independent — and potentially more precise — than calculations using current methods. "From the point of view of measuring the Hubble constant, that's our gold-plated source", says Holz.

The LIGO and Virgo teams estimate that they have a 90% chance of finding more events in the data that LIGO has already collected. They are confident that by the time the next run finishes, the event count will be at least 5, growing to perhaps 35 by the end of a run scheduled to start in 2017.

"To be honest," says Holz, "I find it really hard to believe that the Universe is really doing this stuff. But it's not science fiction. It really happened." ■ SEE EDITORIAL P.413

**Davide Castelvecchi** *is a reporter for* Nature *in London.*

1. Abbott, B. P. *et al. Phys. Rev. Lett.* **116,** 061102 (2016).
2. Abbott, B. P. *et al. Astrophys. J. Lett.* **818,** L22 (2016).
3. Portegies Zwart, S. F., Baumgardt, H., Hut, P., Makino, J. & McMillan, S. L. W. *Nature* **428,** 724–726 (2004).
4. Connaughton, V. *et al.* Preprint at http://arxiv.org/abs/1602.03920 (2016).

Moon jellyfish (*Aurelia aurita*) contain more calories than some other jellyfish.

# *The secret lives of* JELLYFISH

**Long regarded as minor players in ocean ecology, jellyfish are actually important parts of the marine food web.**

BY GARRY HAMILTON

Jennifer Purcell watches intently as the boom of the research ship *Skookum* slowly eases a 3-metre-long plankton net out of Puget Sound near Olympia, Washington. The marine biologist sports a rain suit, which seems odd for a sunny day in August until the bottom of the net is manoeuvred in her direction, its mesh straining from a load of moon jellyfish (*Aurelia aurita*). Slime drips from the bulging net, and long tentacles dangle like a scene from an alien horror film. But it does not bother Purcell, a researcher at Western Washington University's marine centre in Anacortes. Pushing up her sleeves, she plunges in her hands and begins to count and measure the messy haul with an assuredness borne from nearly 40 years studying these animals.

JEFF WILDERMUTH/NATL GEOGR. CREATIVE

Most marine scientists do not share her enthusiasm for the creatures. Purcell has spent much of her career locked in a battle to find funding and to convince ocean researchers that jellyfish deserve attention. But she hasn't had much luck. One problem is the challenges that come with trying to study organisms that are more than 95% water and get ripped apart in the nets typically used to collect other marine animals. On top of that, outside the small community of jellyfish researchers, many biologists regard the creatures as a dead end in the food web — sacs of salty water that provide almost no nutrients for predators except specialized ones such as leatherback sea turtles (*Dermochelys coriacea*), which are adapted to consume jellies in large quantities.

"It's been very, very hard to convince fisheries scientists that jellies are important," says Purcell.

But that's starting to change. Among the crew today are two fish biologists from the US National Oceanic and Atmospheric Administration (NOAA) whose research had previously focused on the region's rich salmon stocks. A few years ago, they discovered that salmon prey such as herring and smelt tend to congregate in different areas of the sound from jellyfish[1] and they are now trying to understand the ecological factors at work and how they might be affecting stocks of valuable fish species. But first, the researchers need to know how many jellyfish are out there. For this, the team is taking a multipronged approach. They use a seaplane to record the number and location of jellyfish aggregations, or 'smacks', scattered about the sound. And on the research ship, a plankton net has been fitted with an underwater camera to reveal how deep the smacks reach.

Correigh Greene, one of the NOAA scientists on board, says that if salmon populations are affected in some way by jellyfish, "then we need to be tracking them".

From the fjords of Norway to the vast open ocean waters of the South Pacific, researchers are taking advantage of new tools and growing concern about marine health to probe more deeply into the roles that jellyfish and other soft-bodied creatures have in the oceans. Initially this was driven by reports of unusually large jellyfish blooms wreaking havoc in Asia, Europe and elsewhere, which triggered fears that jellyfish were taking over the oceans. But mounting evidence is starting to convince some marine ecologists that gelatinous organisms are not as irrelevant as previously presumed.

Some studies show that the animals are important consumers of everything from microscopic zooplankton to small fish, others suggest that jellies have value as prey for a wide range of species, including penguins, lobsters and bluefin tuna. There's also evidence that they might enhance the flow of nutrients and energy between the species that live in the sunlit surface waters and those in the impoverished darkness below.

"We're all busy looking up at the top of the food chain," says Andrew Jeffs, a marine biologist at the University of Auckland in New Zealand. "But it's the stuff that fills the bucket and looks like jelly snot that is actually really important in terms of the planet and the way food chains operate."

### A MASS OF MUSH

The animals in question are descendants of some of Earth's oldest multicellular life forms. The earliest known jellyfish fossil dates to more than 550 million years ago, but some researchers estimate that they may have been around for 700 million years, appearing long before fish.

They're also surprisingly diverse. Some are tiny filter feeders that can prey on the zooplankton that few other animals can exploit. Others are giant predators with bells up to two metres in diameter and tentacles long enough to wrap around a school bus — three times. Jellyfish belong to the phylum Cnidaria and have stinging cells that are potent enough in some species to kill a human. Some researchers use the term jellyfish, or 'jellies' for short, to refer to all of the squishy forms in the ocean. But others prefer the designation of 'gelatinous zooplankton' because it reflects the amazing diversity among these animals that sit in many different phyla: some species are closer on the tree of life to humans than they are to other jellies. Either way, the common classification exists mainly for one dominant shared feature — a body plan that is based largely on water.

This structure can make gelatinous organisms hard to see. Many are also inaccessible, living far out at sea or deep below the light zone. They often live in scattered aggregations that are prone to dramatic population swings, making them difficult to census. Lacking hard parts, they're extremely fragile.

"It's hard to find jellyfish in the guts of predators," says Purcell. "They're digested very fast and they turn to mush soon after they're eaten."

For most marine biologists, running into a mass of jellyfish is nothing but trouble because their collection nets get choked with slime. "It's not just that we overlooked them," says Jonathan Houghton at Queen's University Belfast, UK. "We actively avoided them."

> "IT'S NOT JUST THAT WE OVERLOOKED THEM. WE ACTIVELY AVOIDED THEM."

But over the past decade and a half, jellyfish have become increasingly difficult to ignore. Enormous blooms along the Mediterranean coast, a frequent summer occurrence since 2003, have forced beaches to close and left thousands of bathers nursing painful stings. In 2007, venomous jellyfish drifted into a salmon farm in Northern Ireland, killing its entire stock of 100,000 fish. On several occasions, nuclear power plants have temporarily shut down operations owing to jelly-clogged intake pipes.

### JELLIES ON THE RAMPAGE

The news spurred scientists to take a closer look at the creatures. Marine biologist Luis Cardona at the University of Barcelona in Spain had been studying mostly sea turtles and sea lions. But around 2006, he shifted some of his attention to jellyfish after large summer blooms of mauve stingers (*Pelagia noctiluca*) had become a recurring problem for Spain's beach-goers. Cardona was particularly concerned by speculation that the jellyfish were on the rampage because overfishing had reduced the number of predators. "That idea didn't have very good scientific support," he says. "But it was what people and politicians were basing their decisions on, so I decided to look into it."

For this he turned to stable-isotope analysis, a technique that uses the chemical fingerprint of carbon and nitrogen in the tissue of animals to tell what they have eaten. When Cardona's team analysed 20 species of predator and 13 potential prey, it was surprised to find that jellies had a major role in the diets of bluefin tuna (*Thunnus thynnus*), little tunny (*Euthynnus alletteratus*) and spearfish (*Tetrapturus belone*)[2]. In the case of juvenile bluefins, jellyfish and other gelatinous animals represented up to 80% of the total food intake. "According to our models they are probably one of the most important prey for juvenile bluefin tuna," says Cardona.

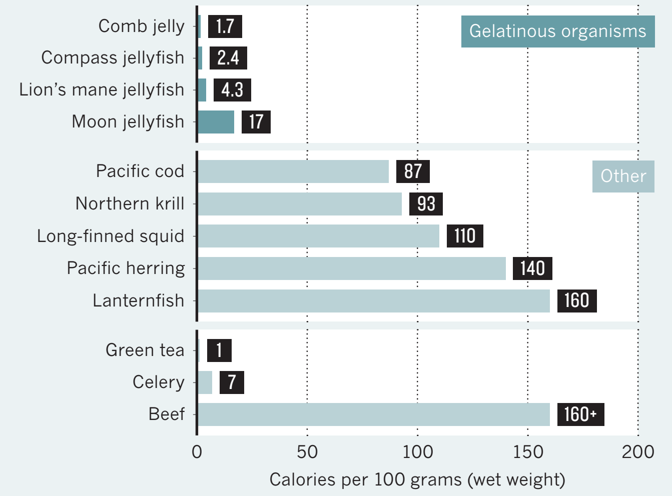
Some researchers have challenged the findings, arguing that stable-isotope results can't always distinguish between prey that have similar diets — jellyfish and krill both eat phytoplankton, for instance. "I'm sure it's not true," Purcell says of the diet analysis. Fast-moving fish, she says, "have the highest energy requirements of anything that's out there. They need fish to eat — something high quality, high calorie."

But Cardona stands by the results, pointing out that stomach-content analyses on fish such as tuna have found jellyfish, but not krill. What's more, he conducted a different diet study[3] that used fatty acids as a signature, which supported his earlier results on jellyfish, he says. "They're probably playing a more relevant role in the pelagic ecosystem of the western Mediterranean than we originally thought."

Researchers are reaching the same conclusion elsewhere in the world. On an expedition to Antarctica in 2010–11, molecular ecologist Simon Jarman gathered nearly 400 scat samples to get a better picture of the diet of Adélie penguins (*Pygoscelis adeliae*), a species thought to be threatened by global warming. Jarman, who works at the Australian Antarctic Division in Kingston, reported in 2013 that DNA analysis of the samples revealed that jellyfish are a common part of the penguin's diet[4]. Work

## LEAN CUISINE

Compared with other marine prey, jellyfish and other gelatinous creatures provide relatively few calories by weight. They are like the green tea of the sea.



SOURCES: J. SPITZ *ET AL. J. MAR. SCI.* **67,** 909–915 (2010); T. K. DOYLE *ET AL. J. EXP. MAR. BIOL. ECOL.* **343,** 239–252 (2007)

"IT'S OVERTURNING THE PARADIGM THAT JELLYFISH ARE DEAD ENDS IN THE FOOD WEB."

that has yet to be published suggests the same is true for other Southern Ocean seabirds.

"Albatrosses, gentoo penguins, king penguins, macaroni and rockhopper penguins — all of them eat jellyfish to some extent," says Jarman (see 'Lean cuisine'). "Even though jellyfish may not be the most calorifically important food source in any area, they're everywhere in the ocean and they're contributing something to many top-level predators."

And some parts of jellyfish hold more calories than others. Fish have been observed eating only the gonads of reproductive-stage jellyfish, suggesting a knack for zeroing in on the most energy-rich tissues.

Through DNA analyses, researchers are also discovering more about how jellyfish function as refuges in the open ocean. Scientists have long known that small fish, crustaceans and a wide range of other animals latch on to jellyfish to get free rides. But in the past few years, it has become clear that the hitchhikers also dine on their transport.

In the deep waters of the South Pacific and Indian oceans, Jeffs has been studying the elusive early life stages of the spiny lobster (*Panulirus cygnus*). During a 2011 plankton-collecting expedition 350 kilometres off the coast of Western Australia, he and his fellow researchers hauled in a large salp (*Thetys vagina*), a common barrel-shaped gelatinous animal. The catch also included dozens of lobster larvae, including six that were embedded in the salp itself. DNA analysis of the lobsters' stomach glands revealed that the larvae had been feeding on their hosts[5].

Jeffs now suspects that these crustaceans, which support a global fishery worth around US$2 billion a year, depend heavily on this relationship. "What makes the larvae so successful in the open ocean," he says, "is that they can cling to what is basically a big piece of floating meat, like a jellyfish or a big salp, and feed on it for a couple of weeks without exerting any energy at all."

### WHERE DID THEY GO?

Researchers are starting to recognize that jellyfish are important for other reasons, such as transferring nutrients from one part of the ocean to another. Biological oceanographer Andrew Sweetman at the International Research Institute of Stavanger in Norway has seen this in his studies of 'jelly falls', a term coined to describe what happens when blooms crash and a large number of dead jellies sink rapidly to the sea floor.

In November 2010, Sweetman began to periodically lower a camera rig 400 metres to the bottom of Lurefjorden in southwestern Norway to track the fate of this fjord's dense population of jellyfish[6]. Previous observations from elsewhere had suggested that dead jellies pile up and rot, lowering oxygen levels and creating toxic conditions. But Sweetman was surprised to find almost no dead jellies on the sea floor. "It didn't make sense."

He worked out what was happening in 2012, when he returned to the fjord and lowered traps baited with dead jellyfish and rigged with video cameras. The footage from the bottom of the fjord showed scavengers rapidly consuming the jellies. "We had just assumed that nothing was going to be eating them," he says.

Back on land, Sweetman calculated[7] that jelly falls increased the amount of nitrogen reaching the bottom by as much as 160%. That energy is going back into the food web instead of getting lost through decay, as researchers had thought. He's since found similar results using remotely operated vehicles at much greater depths in remote parts of the Pacific Ocean. "It's overturning the paradigm that jellyfish are dead ends in the food web," says Sweetman.

Such discoveries have elicited mixed responses. For Richard Brodeur, a NOAA fisheries biologist based in Newport, Oregon, the latest findings do not change the fact that fish and tiny crustaceans such as krill are the main nutrient source for most of the species that are valued by humans. If jellyfish are important, he argues, it is in the impact they can have as competitors and predators when their numbers get out of control. In one of his current studies, he's found that commercially valuable salmon species such as coho (*Oncorhynchus kisutch*)and Chinook (*Oncorhynchus tshawytscha*) that are caught where jellyfish are abundant have less food in their stomachs compared with those taken from where jellies are rare, suggesting that jellyfish may have negative impacts on key fish species. "If you want fish resources," he says, "having a lot of jellyfish is probably not going to help."

But other researchers see the latest findings as reason to temper the growing vilification of jellyfish. In a 2013 book chapter[8], Houghton and his three co-authors emphasized the positive side of jellies in response to what they saw as "the flippant manner in which wholesale removal of jellyfish from marine systems is discussed". As scientists gather more data, they hope to get a better sense of exactly what role jellyfish have in various ocean regions. If jellies turn out to be as important as some data now suggest, the population spikes that have made the headlines in the past decade could have much wider repercussions than previously imagined.

Back in Puget Sound, Greene is using a camera installed on a net to gather census data on a jellyfish smack. He watches video from the netcam as it slowly descends through a dense mass of creamy white spheres. At a depth of around 10 metres, the jelly curtain finally begins to thin out. Later, Greene makes a crude estimate. "Two point five to three million," he says, before adding after a brief pause, "that's a lot of jellyfish."

A more careful count will come later. Right now there's plenty of slime to be hosed off the back deck. Once that's taken care of, the ship's engines come to life. The next jellyfish patch awaits. ■

**Garry Hamilton** *is a freelance writer in Seattle, Washington.*

1. Rice, C. A., Duda, J. J., Greene, C. M. & Karr, J. R. *Mar. Coast. Fish.* **4,** 117–128 (2012).
2. Cardona, L., Álvarez de Quevedo, I., Borrell, A. & Aguilar, A. *PLoS ONE* **7,** e31329 (2012).
3. Cardona, L., Martínez-Iñigo, L., Mateo, R. & González-Solís, J. *Mar. Ecol. Prog. Ser.* **531,** 1–14 (2015).
4. Jarman, S. N. *et al. PLoS ONE* **8,** e82227 (2013).
5. O'Rorke, R. *et al. ICES J. Mar. Sci.* **72** (Suppl. 1), i124–i127 (2014).
6. Sweetman, A. K., Smith, C. R., Dale, T. & Jones, D. O. B. *Proc. R. Soc. B* **281,** 20142210 (2014).
7. Sweetman, A. K. & Chapman, A. *Front. Mar. Sci.* **2,** 47 (2015).
8. Doyle, T. K., Hays, G. C., Harrod, C. & Houghton, J. D. R. in *Jellyfish Blooms* 105–127 (Springer, 2013).

# COMMENT

UMICORE

Workers at Umicore in Brussels separate out precious metals from electronic waste.

# Circular economy

A new relationship with our goods and materials would save resources and energy and create local jobs, explains **Walter R. Stahel**.

When my battered 1969 Toyota car approached the age of 30, I decided that her body deserved to be remanufactured. After 2 months and 100 hours of work, she returned home in her original beauty. "I am so glad you finally bought a new car," my neighbour remarked. Quality is still associated with newness not with caring; long-term use as undesirable, not resourceful.

Cycles, such as of water and nutrients, abound in nature — discards become resources for others. Yet humans continue to 'make, use, dispose'. One-third of plastic waste globally is not collected or managed[1].

There is an alternative. A 'circular economy' would turn goods that are at the end of their service life into resources for others, closing loops in industrial ecosystems and minimizing waste (see 'Closing loops'). It would change economic logic because it replaces production with sufficiency: reuse what you can, recycle what cannot be reused, repair what is broken, remanufacture what cannot be repaired. A study of seven European nations found that a shift to a circular economy would reduce each nation's greenhouse-gas emissions by up to 70% and grow its workforce by about 4% — the ultimate low-carbon economy (see go.nature.com/biecsc).

The concept grew out of the idea of substituting manpower for energy, first described 40 years ago in a report[2] to the European Commission by me and Geneviève Reday-Mulvey while we were at the Battelle Research Centre in Geneva, Switzerland. The early 1970s saw rising energy prices and high unemployment. As an architect, I knew that it took more labour and fewer resources to refurbish buildings than to erect new ones. The principle is true for any stock or capital, from mobile phones to arable land and cultural heritage.

Circular-economy business models fall in two groups: those that foster reuse and extend service life through repair, remanufacture, upgrades and retrofits; and those that turn old goods into as-new resources by recycling the materials. People — of all ages and skills — are central to the model. Ownership gives way to stewardship; consumers become users and creators[3]. The remanufacturing and repair of old goods, buildings and infrastructure creates skilled jobs in local workshops. The experiences of ▶

workers from the past are instrumental.

Yet a lack of familiarity and fear of the unknown mean that the circular-economy idea has been slow to gain traction. As a holistic concept, it collides with the silo structures of academia, companies and administrations. For economists who work with gross domestic product (GDP), creating wealth by making things last is the opposite of what they learned in school. GDP measures a financial flow over a period of time; circular economy preserves physical stocks. But concerns over resource security, ethics and safety as well as greenhouse-gas reductions are shifting our approach to seeing materials as assets to be preserved, rather than continually consumed.

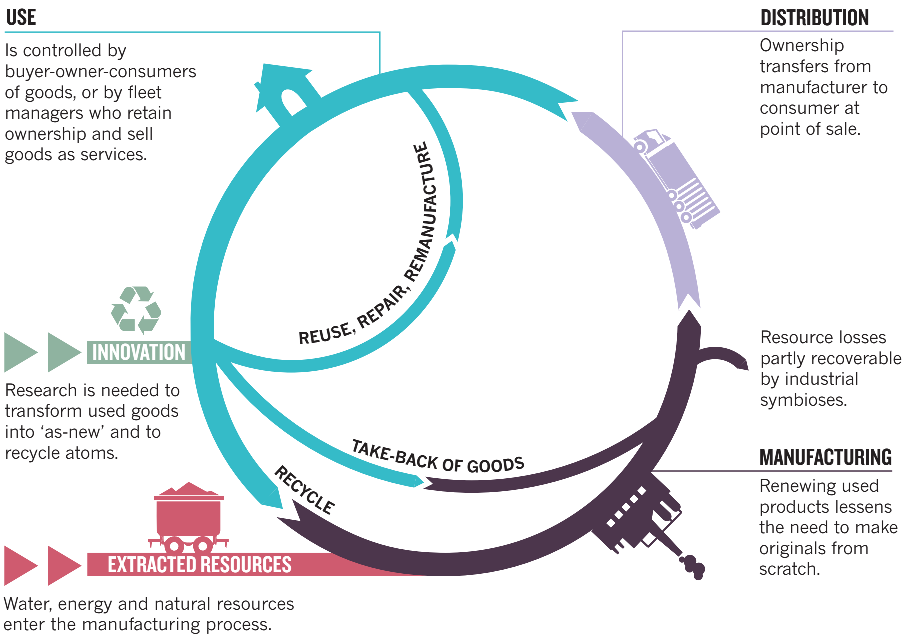
In the past decade, South Korea, China and the United States have started research programmes to foster circular economies by boosting remanufacturing and reuse. Europe is taking baby steps. The Swedish Foundation for Strategic Environmental Research (Mistra) and the EU Horizon 2020 programme published their first call for circular-economy proposals in 2014. The European Commission submitted a Circular Economy Package to the European Parliament last December. Since 2010, the Ellen MacArthur Foundation, founded by the round-the-world yachtswoman, has been boosting awareness of the idea in manufacturers and policymakers. And circular-economy concepts have been successfully applied on small scales since the 1990s in eco-industrial parks such as the Kalundborg Symbiosis in Denmark, and in companies that include Xerox (selling modular goods as services), Caterpillar (remanufacturing used diesel engines) and USM Modular Furniture. Selling services rather than goods is familiar in hotels and in public transport; it needs to become mainstream in the consumer realm.

Few researchers are taking note. Excellence in metallurgical and chemical sciences is a precondition for a circular economy to succeed. Yet there is too little research on finding ways to disassemble material blends at the atomic level. The body of a modern car incorporates more than a dozen steel and aluminium alloys, each of which needs to be retrieved.

Circular-economy knowledge is concentrated in big industries and dispersed across small–medium enterprises (SMEs). It must be brought into academic and vocational training. A broad 'bottom up' movement will emerge only if SMEs can hire graduates who have the economic and technical know-how to change business models. Governments and regulators should adapt policy levers, including taxation, to promote a circular economy in industry. And scientists should scan the horizon for innovations that could be patented and licensed to pave the way for greater leaps in splitting up molecules to recycle atoms.

## CLOSING LOOPS

Using resources for the longest time possible could cut some nations' emissions by up to 70%, increase their workforces by 4% and greatly lessen waste.



**USE**
Is controlled by buyer-owner-consumers of goods, or by fleet managers who retain ownership and sell goods as services.

**DISTRIBUTION**
Ownership transfers from manufacturer to consumer at point of sale.

Resource losses partly recoverable by industrial symbioses.

**MANUFACTURING**
Renewing used products lessens the need to make originals from scratch.

**INNOVATION**
Research is needed to transform used goods into 'as-new' and to recycle atoms.

**EXTRACTED RESOURCES**
Water, energy and natural resources enter the manufacturing process.

ADAPTED FROM KNOWLEDGE TRANSFER NETWORK

## SYSTEMS THINKING

There are three kinds of industrial economy: linear, circular and performance.

A linear economy flows like a river, turning natural resources into base materials and products for sale through a series of value-adding steps. At the point of sale, ownership and liability for risks and waste pass to the buyer (who is now owner and user). The owner decides whether old tyres will be reused or recycled — as sandals, ropes or bumpers — or dumped. The linear economy is driven by 'bigger-better-faster-safer' syndrome — in other words, fashion, emotion and progress. It is efficient at overcoming scarcity, but profligate at using resources in often-saturated markets. Companies make money by selling high volumes of cheap and sexy goods.

A circular economy is like a lake. The reprocessing of goods and materials generates jobs and saves energy while reducing resource consumption and waste. Cleaning a glass bottle and using it again is faster and cheaper than recycling the glass or making a new bottle from minerals. Vehicle owners can decide whether to have their used tyres repaired or regrooved or whether to buy new or retreaded replacements — if such services exist. Rather than being dumped, used tyres are collected by waste managers and sold to the highest bidder.

THE CIRCULAR ECONOMY
A *Nature* special issue
nature.com/thecirculareconomy

A performance economy goes a step further by selling goods (or molecules) as services through rent, lease and share business models[4,5]. The manufacturer retains ownership of the product and its embodied resources and thus carries the responsibility for the costs of risks and waste. In addition to design and reuse, the performance economy focuses on solutions instead of products, and makes its profits from sufficiency, such as waste prevention.

For example, Michelin has since 2007 sold tyre use 'by the mile' to operators of vehicle fleets. The company has developed mobile workshops to repair and regroove tyres at clients' premises and aims to develop products with longer service lives. Worn tyres are sent to Michelin's regional plants for retreading and reuse. The Swiss company Elite uses the same strategy for hotel mattresses, and textile-leasing companies offer uniforms, hotel and hospital textiles and industrial wipes as a service.

Conventional waste management is driven by minimizing the costs of collection and disposal — landfill versus recycling or incineration. In a circular economy, the objective is to maximize value at each point in a product's life. New jobs will be created and systems are needed at each step.

Commercial markets and collection points are needed for users and manufacturers to take back, bring back or buy back discarded garments, bottles, furniture, computer equipment and building components. Goods that can be reused may be cleaned and re-marketed; recyclables are dismantled and the

Autolib car-sharing schemes free users from the demands of ownership.

GONZALO FUENTES/REUTERS

parts are classified according to their residual value. Worn parts are sold for remanufacturing, broken ones for recycling. These markets used to be common — milk and beer bottles and old iron were once collected regularly from homes. Some have re-emerged as digital global market places, such as eBay.

Professional marketplaces (perhaps online) also need to be set up for the exchange of used parts, such as electric motors, bearings and microchips. Even components of liquid waste, such as lubrication and cooking oils or phosphorus from sewage, can be refined and resold. Scientists should re-market rather than dump their used kit.

Stewardship rules are needed for used goods. Austria is a world leader in this area. Collecting and reusing 'waste' are labour intensive and expensive, but they have been fostered in the nation through taxation changes and by recouping costs through re-marketing rather than scrapping parts.

The ultimate goal is to recycle atoms. This is already done for some metals. The Brussels-based company Umicore extracts gold and copper from electronic waste. The Swiss firm Batrec removes zinc and ferromanganese from batteries. These processes are energy-intensive and recover the metals only partly. To close the recovery loop we will need new technologies to de-polymerize, de-alloy, de-laminate, de-vulcanize and de-coat materials.

Methods and equipment are needed to deconstruct infrastructure and high-rise buildings. For example, the ANA InterContinental hotel in Tokyo was demolished in 2014 beneath a 'turban' that was lowered hydraulically floor by floor to minimize noise and dust emissions. A vertical shaft with a goods lift in the middle of the building allowed the deconstructors to recover components and sort materials while using the lift as a generator.

Services liberate users from the burden of ownership and maintenance and give them flexibility. Examples include: 'power by the hour' for jet and gas turbines; bike and car rentals; laundromats and machine-hire shops. Fleet managers benefit from resource security — the goods of today become the resources of tomorrow at yesterday's prices. Covering the costs of risk and waste within the price of use or hire provides economic incentives to prevent loss and waste over the lifetimes of systems and products.

> ***"We will need new technologies to de-polymerize, de-alloy, de-laminate, de-vulcanize and de-coat materials."***

### SOCIETAL TREND

The circular economy is part of a trend towards intelligent decentralization — witness 3D printing, mass customization of manufacturing, 'labs-on-a-chip' in chemistry and functional services. The French car-sharing service Autolib offers people flexible, hassle-free urban mobility by using small electric cars that have low maintenance costs and can be recharged in reserved parking spaces throughout Paris. Such business models jeopardize the fundamentals of the linear economy — ownership, fashion and emotion — and raise fears in competing companies. For example, car manufacturers' strengths of mass production, patented technologies in combustion engines and gearboxes, big investments in robotic factories and global supply and marketing chains are of little use when competing with local Autolib services.

Public procurement can exploit the potential of the performance economy. Yet despite some successes, governments remain hesitant. NASA decided a decade ago to buy space transport services, leading to start-up companies such as SpaceX competing for contracts using innovative, cheap and reusable equipment. Assigning maintenance costs to the private constructor of the Millau Viaduct in the south of France led the tenderer, Eiffage Construction, to develop a structure that could be erected quickly and would have minimal maintenance and liability costs over its 75-year service life.

### TIPPING POINTS

Realizing a circular economy will take concerted action on several fronts.

Research and innovation are needed at all levels — social, technological and commercial. Economists and environmental and materials scientists need to assess the ecological impacts and costs and benefits of products. Designing products for reuse needs to become the norm, making use of modular systems and standardized components, for instance[6]. More research is needed to convince businesses and governments that a circular economy is feasible.

Communication and information strategies are needed to raise the awareness of manufacturers and the public about their responsibility for products throughout their service lives. For instance, it should be fashion magazines, not science journals, that bang the drum about jewellery sharing, leased jeans and rental designer handbags.

Policymakers should use 'resource-miser' indicators such as value-per-weight and labour-input-per-weight ratios rather than GDP. Policies should focus on performance, not hardware; internalization of external costs, such as emissions and pollution, should be rewarded; stewardship should overrule ownership and its right to destroy. The Internet of Things (in which everyday objects are digitally connected) and Industry 4.0 (intelligent technical systems for mass production) will boost such a shift, but also demand a policy review that considers questions of ownership and liability of data and goods[7,8].

Policies[9] should promote activities that are desired by society and punish those that are not. Taxes should be raised on the consumption of non-renewable resources, not on renewable resources including human labour. Value-added tax (VAT) should be levied on value-added activities, such as mining, construction and manufacturing, but not on value-preserving stock management activities such as reuse, repair and remanufacture. Carbon credits should be given to emissions prevention at the same rate as to reduction.

Societal wealth and well-being should be measured in stock instead of flow, in capital instead of sales. Growth then corresponds to a rise in the quality and quantity of all stocks — natural, cultural, human and manufactured. For example, sustainable forestry management augments natural capital, deforestation destroys it; recovering phosphorus or metals from waste streams maintains natural capital, but dumping it increases pollution; retrofitting buildings reduces energy consumption and increases the quality of built stock[10].

Marrying the three types of economy is a formidable challenge. A shift in policy focus from protecting the environment to promoting business models that are based on full ownership and liability, and that are unlimited in time, rather than imposing a two-year warranty for manufacturing quality, could transform a nation's competitiveness. ■

**Walter R. Stahel** *is founder and director of the Product-Life Institute in Geneva, Switzerland. He is also a member of the Club of Rome and a visiting professor at the Faculty of Engineering and Physical Sciences, University of Surrey, UK.*
*e-mail: wrstahel2014@gmail.com*

1. Ellen MacArthur Foundation, World Economic Forum and McKinsey & Company. *The New Plastics Economy: Rethinking the Future of Plastics* (Ellen MacArthur Foundation, 2016).
2. Stahel, W. R. & Reday-Mulvey, G. *Jobs for Tomorrow: The Potential for Substituting Manpower for Energy* ((Vantage Press, 1981).
3. Stahel, W. R. in *The Circular Economy — A Wealth of Flows* (ed. Webster, K.) 86–103 (Ellen MacArthur Foundation, 2015).
4. Stahel, W. R. *The Performance Economy* (Palgrave, 2006).
5. Stahel, W. R. in *Handbook of Performability Engineering* (ed. Misra, K. B.) Ch. 10, 127–138 (Springer, 2008).
6. Stahel, W. R. in *Our Fragile World: Challenges and Opportunities for Sustainable Development* Vol. II (ed. Tolba, M. K.) Ch. 30, 1553–1568 (UNESCO/EOLSS, 2001).
7. Giarini, O. & Stahel, W. R. *The Limits to Certainty, Facing Risks in the New Service Economy* (Kluwer, 1989).
8. Stahel, W. R. in *The Industrial Green Game: Implications for Environmental Design and Management* (ed. Richards, D. J.) Ch. 4, 91–100 (National Academy Press, 1997).
9. Stahel, W. R. *Phil. Trans. R. Soc. A* **371,** 20110567 (2013).
10. Stahel, W. R. & Clift, R. in *Taking Stock of Industrial Ecology* (eds Clift, R. & Druckman, A.) Ch. 7, 137–158 (Springer, 2016).

**Stalls known as mtumbas ('second-hand' in Swahili) in Nairobi sell repurposed goods, many from the West.**

KELLYRANCK.COM

# Make recycled goods covetable

To reduce consumption and waste we must overcome our squeamishness about repurposing pre-owned possessions, says **Bruce Hood**.

Humans are unique in the animal kingdom in their capacity for materialism. We make, use and trade objects for their symbolic value as much as their functionality. One of the earliest examples of such artefacts— a piece of carved ochre found in the Blombos Cave in South Africa — dates from at least 70,000 years ago. Possessions are extensions of our selves. Beyond making tools, we adorn ourselves and bury our dead with objects.

Objects have social significance. Through them we signal our identity and status to others. Marketing experts know that belongings convey aspirations that owners wish to display to others. Designer goods have cachet because of their expense or exclusivity. To all but the most ascetic among us, it is important to some degree what others think about our choice of gadgets, car, décor or clothing.

These mores of ownership inform the value that we assign fakes or those who own them. When it comes to second-hand goods, most of us care about who previously handled them and what they were used for — we would rather wear the clothing of a beloved celebrity than a murderer. We reverently hand down great-grandma's costume jewellery to the next generation, but toss last season's bling from

the high street. It is as if something tangible has imbued the very substance of the object.

Underpinning these unconscious, often irrational, preferences is psychological essentialism — a belief that identity is conferred by a metaphysical dimension, an essence that cannot be removed, filtered, eradicated or repurposed by physical means. Countering these biases with logic is difficult.

But countered they must be. Essentialism presents a formidable obstacle to accepting — as we must — that all materials can and should be reused or recycled. To realize a circular economy — in which resources are kept in use for as long as possible — the perceived status and value of reused materials must be changed. How? I think the answer requires us to shift from valuing objects on the basis of exclusivity to a valuation that prioritizes historical reuse.

## STATUS SYMBOLS

Some have argued that today's rampant consumerism reflects an obsession with gaining status that originated from our evolved capacity to live in hierarchical social groups in which possessions were equated with success[1]. Status determines reproductive success in many social animals. Just as the male peacock's large and lustrous tail signals his health and strength to prospective mates, so too does evident material wealth in humans.

More than a century ago, US economist and sociologist Thorstein Veblen coined the term conspicuous consumption as the attaining and exhibiting of costly items to impress others[2]. He argued that many people in power, from the Egyptian Pharaohs to the maharajahs of India, flaunted their wealth to signal superiority. Little has changed over millennia.

But our ability to make more possessions has changed. The accumulated store of manufactured goods has risen exponentially with the power of technology to increase production. For example, between 1860 and 1920, US production increased 12–14 times, whereas the population only tripled. The amount of stuff we could make outstripped demand, which needed stimulating to maintain economic growth. Marketing strategies since have reinforced consumerism as a necessary component of self-worth, creating problems from mild binges of 'retail therapy' to pathological over-spending.

This incentive to own does not require much effort — even children are selfish about possessions. More than 80% of preschoolers' conflicts with peers revolve around ownership[3]. Toys are more coveted when they have been touched or named by another child. We soon learn to define ourselves by what we own. Psychologist Sam Gosling, author of *Snoop: What Your Stuff Says About You*[4], has demonstrated links between different personality types and the sorts of objects that adults adorn their personal spaces with as an expression of self-identity. For example, men tend to display trophies and women are more likely to decorate their spaces with objects associated with their relationships.

Because we tend to view ourselves positively, we project greater value onto our own possessions than others would — an impulse called the endowment effect. This bias varies among cultures and is stronger in individualistic compared with interdependent societies. For example, in a 2010 study[5], US adults of European heritage asked for a much higher selling price for their coffee mug compared with Asian American adults. In the same experiment, priming Chinese and Japanese adults to think about themselves shifted the endowment effect in a direction more typical of Westerners.

> **"The psychological bias to value exclusivity undermines the principles of recycling."**

Surprisingly, the endowment effect may be stronger where there are more rather than fewer possessions. For example, the Hadza hunter-gatherers of Tanzania are all equally poor and do not normally overvalue their own possessions. It is the gap between the haves and have-nots that drives possessiveness, it seems. According to Nicholas Christakis, a sociologist at Yale University in New Haven, Connecticut, the endowment effect arises when inequality in individualistic societies is visible to all[6]. When it comes to economic harmony, ignorance — or greater equality — is bliss.

The 'extended self' hypothesis[7] includes in our 'self' everything that we can claim ownership over. A person who owns a nice home, a new car, good furniture and the latest appliances, is recognized by others as someone who has passed the test of personhood in Western society.

The pleasure one derives from a Rolex watch or an Armani suit is largely psychological and is based on perceived desirability rather than on sensory or functional pay-off. Designer brands are esteemed beyond their quality. By definition, a luxury item (the word coming from the Latin *luxus*, meaning excess) generates value from its exclusivity. Lobsters and oysters command high prices today, but in the eighteenth century, before refrigeration allowed them to be shipped to cities, they were the food of poor fishing communities.

Authenticity also matters. Reproduced items or fake brands are valued less, even though they can be indistinguishable from an original. And we cannot always fool ourselves. One study[8] showed that individuals who wore what they believed to be fake designer sunglasses felt sullied and were more inclined to dishonesty, even when the glasses were in fact expensive originals. Even seven-year-olds rate original possessions supposedly belonging to Queen Elizabeth II as more valuable than identical copies[9].

## PSYCHOLOGICAL ESSENTIALISM

The psychological bias to value exclusivity and authenticity undermines the principles of recycling and reuse. Recycled items lack authenticity, which compromises their identity and perceived value.

Most adults reason, for example, that if their gold wedding ring was swapped with a duplicate, it would not be the same ring. If we were told that a small particle of the ring's metal was replaced, we would regard it as the same ring. If we were told that over time the ring was completely renovated, we would still think it the same ring even when there was no original material present — the same as the swapped ring. Thus there is an essential property of the ring beyond its physical make-up that continues its identity.

Such retained identity could operate by contamination. Each new particle of gold added to the ring becomes assimilated into the whole. Simply by touch, objects take on the property of the owner as if by contagion. For example, memorabilia collectors will pay inflated prices for a sweater that they believe was worn by a pop star or famous actor, but much less if it is sterilized. Conversely, they will pay little for one that belonged to a disliked figure (such as a fraudster) unless it has been sterilized[10].

In short, we value old items for their sentimentality, nostalgia or connection with the famous. But not as much as we once did: the Antique Collectors' Club's Annual Furniture Index, based on a mixture of auction and retail prices of 1,400 typical items, has been on the slide since reaching a peak in 2002.

In the same way that conspicuous consumerism was encouraged at the turn of the twentieth century to redress the imbalance between overproduction and demand, policies must now encourage conspicuous non-consumption and reuse as the new signifiers of self-worth.

To address the long-term consequences of unbridled materialism, we need to make having fewer things and using recycled goods more socially desirable. Currently, only a few retailers sell items such as purses and bags that have been ingeniously 'upcycled' from low-value, discarded goods such as cement sacks and tyres. Instead of being niche products, such items should be status symbols. Frugal innovation must become ubiquitous, not just the preserve of poor nations or of times past.

The more recycled material used in an object, the more this quality should be advertised (and rewarded with tax breaks and other market levers). In the same way that food products must declare their constituents and additives, manufactured goods should indicate the extent of their recycled content. Packaging often states the proportion of recycled material used but rarely does the same disclosure appear for the product contained within.

This might start to shift attitudes away from the appeal of the 'brand new' to appreciating the value of the 'brand renewed' — something that will be essential in a sustainable, circular, economy. ■

**Bruce Hood** *is professor of developmental psychology in society at the University of Bristol, UK, and founder of Speakezee.org.*
*e-mail: bruce.hood@bristol.ac.uk*

1. Stake, J. E. *Phil. Trans. R. Soc. Lond. B* **359,** 1763–1774 (2004).
2. Veblen, T. *The Theory of the Leisure Class* (Macmillan, 1899).
3. Hay, D. F. & Ross, H. S. *Child Dev.* **53,** 105–113 (1982).
4. Gosling, S. *Snoop: What Your Stuff Says About You* (Basic, 2009).
5. Maddux, W. W. *et al. Psychol. Sci.* **21,** 1910–1917 (2010).
6. Nishi, A., Shirado, H., Rand, D. G. & Christakis, N. A. *Nature* **526,** 426–429 (2015).
7. Belk, R. W. *J. Consum. Res.* **15,** 139–168 (1988).
8. Gino, F., Norton, M. I. & Ariely, D. *Psychol. Sci.* **21,** 712–720 (2010).
9. Hood, B. & Bloom, P. *Cognition* **106,** 455–462 (2008).
10. Newman, G. E. & Bloom, P. *Proc. Natl Acad. Sci. USA* **111,** 3705–3708 (2014).

The Suzhou New District was one of the first industrial parks in China's circular-economy programme.

YONGSKY/DREAMSTIME.COM

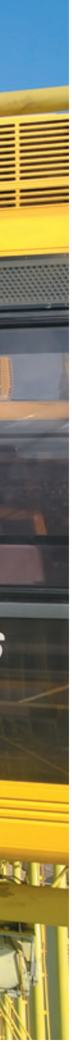
# Lessons from China

## The country consumes the most resources in the world and produces the most waste — but it also has the most advanced solutions, say **John A. Mathews** and **Hao Tan**.

China's consumption of the world's resources is reaching crisis levels. To produce 46% of global aluminium, 50% of steel and 60% of the world's cement[1] in 2011, it consumed more raw materials than the 34 countries of the Organisation for Economic Co-operation and Development (OECD) combined: 25.2 billion tonnes.

The nation's resource use is inefficient. China requires 2.5 kilograms of materials to generate US$1 of gross domestic product (GDP) compared with 0.54 kilograms in OECD countries (in 2005 dollars, adjusted for purchasing power parity). And it is wasteful. In 2014, China generated 3.2 billion tonnes of industrial solid waste, only 2 billion tonnes of which was recovered by recycling, composting, incineration or reuse. By comparison, firms and households in the 28 countries of the European Union generated 2.5 billion tonnes of waste in 2012, of which 1 billion was recycled or used for energy. In 2025, China is expected to produce almost one-quarter of the world's municipal solid waste[2].

Unchecked, such levels of consumption and waste will strain the nation and the planet. In December 2015, a landslide at a waste dump in Shenzhen killed 73 people. China has also seen an increasing number of protests by local residents over waste-incineration projects in recent years. The



geopolitical costs could soar as China becomes more dependent on imported resources from unstable parts of the world. Fuels and minerals accounted for 30% of the total cost of China's imports in 2012, compared with just over 5% in 1990.

The country is taking action. For the past decade, China has led the world in promoting the recirculation of waste materials through setting targets and adopting policies, financial measures and legislation. The ultimate goal is a 'circular economy' — closing industrial loops to turn outputs from one manufacturer into inputs for another. This approach reduces the consumption of virgin materials and the generation of waste.

Progress has been modest and the obstacles to transforming the economy are formidable. Western countries have struggled for decades to get companies to collaborate along a supply chain. China has the advantage that more than half of its manufacturing activities are conducted in industrial parks and export processing zones. Targeting these parks is beginning to slash the intensity of China's resource use.

For example, the Suzhou New District is a 52-square-kilometre region for technological and industrial development near Shanghai, where around 4,000 manufacturing firms operate. There, manufacturers of printed circuit boards use copper that is recovered from waste from elsewhere in the park, rather than using virgin copper produced by mining firms[3].

No other country has such ambitions. Germany and Japan have comprehensive plans for recycling (through Germany's Closed Substance Cycle and Waste Management Act of 1996 and Japan's 2000 Fundamental Law for Establishing a Sound Material-cycle Society). The European Commission announced a Circular Economy Package in December 2015 but has yet to implement it.

The United States has hundreds of corporate recycling initiatives (including those of the machinery company Caterpillar and Interface, a carpet manufacturer). The United States also has a handful of regional programmes such as the Zero Waste scheme in San Francisco, California. Other initiatives involving closing loops to attain 'industrial symbiosis'[4] — in which waste products of one firm become the raw materials of another — are in place in Yokohama, Japan; in Ulsan, South Korea; and in Kwinana, Australia[5]. All these are limited in their impacts and scale.

## AMBITIOUS PLANS

Chinese interest in the circular economy was piqued in the 1990s[6] by Germany and Japan's recycling laws. In 2005, China's State Council issued a policy paper (see go.nature.com/cnozhg; in Chinese) recognizing the economic and environmental risks of the nation's heavy resource exploitation, and acknowledging the circular economy as the principal means of dealing with them. The country's planning agency, the National Development and Reform Commission (NDRC) and bodies such as the Ministry of Environmental Protection have since developed circular-economy principles and promoted exemplars of industrial symbiosis, such as at the Rizhao Economic and Technology Development Zone[7].

Taxation, fiscal, pricing and industrial policies were introduced. A fund was allocated to support the conversion of industrial parks into eco-industrial agglomerations. Tax breaks were provided to enterprises in the reuse sector. To finance the initiatives through concessionary loans or direct capital financing, the NDRC joined with financial regulators including China's central bank and its banking and securities regulatory commissions.

> ***"China has led the world in promoting the recirculation of waste materials."***

A whole chapter in the country's 11th Five-Year Plan (for 2006–10) was devoted to the circular economy. And a 2008 circular-economy 'promotion law' demanded that local and provincial governments consider such issues in their investment and development strategies. Targets were enacted for the coal, steel, electronics, chemical and petrochemical industries. The circular economy was upgraded to a national development strategy in the 12th Five-Year Plan (2011–15).

Objectives included reusing 72% of industrial solid waste by 2015 and raising resource productivity (economic output per unit resources used) by 15%. The plan laid out a three-pronged '10–100–1,000' strategy: 10 major programmes focusing on recycling industrial wastes, conversion of industrial parks, remanufacturing, urban mining, and the development of waste-collection and recycling systems; 100 demonstration cities such as Suzhou and Guangzhou; and 1,000 demonstration enterprises or industrial parks nationwide. In 2012, the NDRC and the finance ministry called for 50% of national industrial parks and 30% of provincial ones to complete circular-economy transformation initiatives by 2015, with an aim of achieving close to zero discharge of pollutants.

In 2013, the State Council released a national strategy for achieving a circular economy — the first such strategy in the world. Further targets for 2015 included increasing energy productivity (GDP per unit energy) by 18.5% relative to 2010, raising water productivity by 43%, and for the output of the recycling industry to reach 1.8 trillion yuan (US$276 billion) compared with 1 trillion yuan in 2010. Others include reusing at least 75% of coal gangue (worthless rock present in deposits) from coal mining or 70% of pulverized fuel ash, a product of coal combustion, from electricity generation.

Some of these targets have been extended in China's 13th Five-Year Plan, which was published this month.

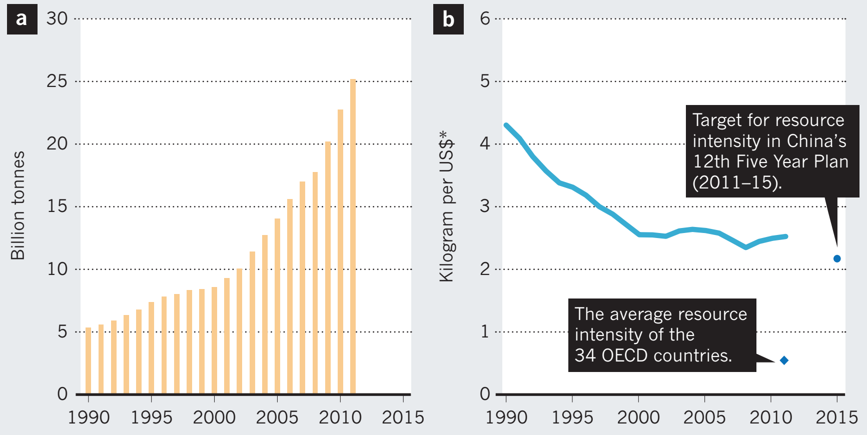
## REPORT CARD

How has China done? Last year, its National Bureau of Statistics analysed[8] progress since 2005 on four measures: resource intensity (resources used per unit GDP), waste intensity (waste per unit GDP), waste recycling rate and pollutant treatment rate.

By 2013, resource intensity and waste intensity had improved by 34.7% and 46.5%, respectively, a clear sign that resource consumption (of metal, water,

## CONSUMPTION PROBLEM

China's economy consumes more and more materials each year (**a**). But as the country becomes more efficient, it uses fewer resources for each dollar of gross domestic product generated (a measure known as resource intensity; **b**).



*2005 dollars, adjusted for purchasing power parity; OECD, Organisation for Economic Co-operation and Development.

SOURCE: OECD

energy and biomass) is decoupling from economic growth in relative terms. The treatment rate of pollution, including sewage, the decontamination of urban residential waste and the reduction of major pollutants, also increased, by 74.6%. The recycling and reuse of waste improved more slowly, by 8.2%. A circular-economy development index created by the statistics bureau aggregating all these indicators grew from 100 in 2005 to 137.6 in 2013.

OECD statistics reveal that China's resource intensity fell: from 4.3 kilograms of materials per unit GDP in 1990 to 2.5 kilograms in 2011. However, China's overall resource consumption rose fivefold during these two decades, from 5.4 billion tonnes to 25.2 billion tonnes, as a result of its economic boom (see 'Consumption problem').

## PARK LIFE

The Suzhou New District (SND) is an exemplar of circular-economy initiatives. In 2005, it was selected as one of the first 13 industrial parks to participate in China's national circular-economy pilot programme. In 2008, it was one of three national eco-industrial park demonstration sites in the country, along with the nearby China–Singapore Suzhou Industrial Park and the Tianjin Economic-Technological Development Area.

The SND is larger than Western examples of industrial symbiosis such as Kalundborg in Denmark, which was the first case of closed-loop recycling since the 1980s. Kalundborg involves a dozen or so firms sharing energy, water, steam and waste-recycling processes. By 2014, the SND hosted more than 16,000 enterprises and almost 4,000 manufacturing firms, many in IT, electronics, biotech and medical-device manufacturing. The total output of the industrial sector of the SND (including manufacturing, mining and utilities) was 288 billion yuan in 2015.

Initiatives set out to plug gaps in chains of industries within industrial parks. For example, the SND administration identified the recycling and recirculation of metal resources such as gold and copper as a gap in the park's printed-circuit-board supply chain. A venture was formed with Dowa Metal in Japan to establish an advanced metal-recycling business in the SND. Waste etching solution that is generated in copper laminating and circuit-board manufacturing in the SND is treated and returned by others based in the park. Electronic-waste companies such as Dowa reclaim the copper and water from the sludge created by circuit-board processing.

In other examples, a producer of kaolin (a type of clay) turns residues from mining into inputs for the production of sulfuric acid and construction materials; a paper manufacturer takes waste ammonia from a chemical company to use for desulfurization in its process; and industrial water recycling is undertaken on site.

According to data from the SND, between 2005 and 2010, the energy intensity of the district dropped by 20% (down to 0.57 tonnes of coal equivalent per 10,000 yuan of GDP, compared with the national 2010 level of 1.24 tonnes of coal equivalent per 10,000 yuan of GDP). During the same period, the park's oxidizable organic pollutants in water dropped by 47%, and emissions of sulfur dioxide by 38% (ref. 9). The utilization rate of industrial solid wastes and the recycling rate of industrial water reached 96% and 91% in 2010, much higher than the national averages (69% and 86%)[10].

The main obstacle is getting firms linked by supply chains to cooperate in turning outputs into inputs — as in the copper-extraction example. Some observers might see China's top-down approach to such issues as problematic, but it is clear that the tradition of managing industrial parks through local institutions and governments is able to cut through the problem by offering rewards to firms that collaborate. Thus the issue is reframed from one involving individual firms to one that involves their collective decisions.

> ***"Some industries lend themselves to circular initiatives more than others."***

The economic benefits are clear. Recycled, regenerated and locally sourced raw materials are usually cheaper, increasing profits. State involvement in the economy turns out to be an advantage, and underpins how progress depends on countries' abilities to implement as well as develop industrial policies.

Some industries lend themselves to circular initiatives more than others. For example, recirculation of metal scrap is straightforward, but extracting metal from industrial sludge is more chemically demanding. China's move away from primary industries to secondary ones, such as solar-panel manufacturing, will reap benefits from the circular economy. And increasing reliance on home-regenerated materials rather than imports will increase the country's resource security.

## NEXT STEPS

China must still do much more. It needs a national goal and road map to achieve a level of resource intensity that is similar to that of OECD countries (currently around 0.5 kilograms per dollar of GDP). And it must champion regional and provincial achievements, giving rewards to eco-industrial parks that perform best. Data should be reported regularly. SND data are five years old, for example. Seeing the fiscal benefits, companies should have incentives to release accurate data.

Primary industries such as iron, steel and aluminium need strong targets for recirculation as part of the thirteenth and subsequent five-year plans. Increasingly, secondary industries such as wind energy, battery production and biotech should be assessed on the basis of their recirculation potential and performance over their whole life cycles.

Better circular-economy metrics need to be developed. The circular-economy index of China's statistics bureau needs clarification on what it means and what it measures. The OECD should similarly draw up reporting guidelines for all countries to follow. Researchers need to collaborate with China to improve metrics and conduct case studies of industrial symbiosis.

Mainstream economics perpetuates linear thinking with concepts such as GDP and the use of GDP growth as a sole performance measure for national economies. Performance measures such as circulation of resources need to be introduced into economists' models, to create an interest in the real flow of resources that underpins abstractions such as income and wealth.

In our view, the only solution to the world's resource-security problem is to move away from the linear economy and embrace the circular economy. China's strategies are a significant step forwards in bridging the global gap between economic and ecological sustainability. ■

**John A. Mathews** *is professor of strategy at the Macquarie Graduate School of Management at Macquarie University, Sydney, Australia.* **Hao Tan** *is senior lecturer in international business at the Newcastle Business School, University of Newcastle, Callaghan, Australia.*
*e-mail: john.mathews@mgsm.edu.au; haotan1@gmail.com*

1. Mathews, J. A. & Tan, H. *China's Renewable Energy Revolution* (Palgrave, 2015).
2. Hoornweg, D., Bhada-Tata, P. & Kennedy, C. *Nature* **502,** 615–617 (2013).
3. Wen, Z. G. & Meng, X. Y. *J. Clean. Prod.* **90,** 211–219 (2015).
4. Chertow, M. *Ann. Rev. Energy Environ.* **25,** 313–337 (2000).
5. Mathews, J. & Tan, H. *J. Indust. Ecol.* **15,** 435–457 (2011).
6. Zhu, D. *Sci. Technol. Rev.* **9,** 39–43 (1998; in Chinese).
7. Yu, F., Han, F. & Cui, Z. *J. Clean. Prod.* **87,** 339–347 (2015).
8. National Bureau of Statistics of the People's Republic of China. *The National Circular Economy Development Index Achieved 137.6 in 2013* (NBS, 2015); available at http://go.nature.com/vft5fz (in Chinese).
9. Xu, Y. *Environ. Prot. Circ. Econ.* **9,** 10–13 (2015; in Chinese).
10. State Council of the People's Republic of China. *Circular Economy Development Strategy and Immediate Plan of Action* (State Council, 2013); available at http://go.nature.com/hmcxiz (in Chinese).

ROLLS ROYCE

The Trent 1000 engine: Rolls-Royce has run a recycling programme for more than a decade.

CIRCULAR ECONOMY

# Getting the circulation going

In linear economics, objects of desire from skyscrapers to paperclips are waste waiting to happen. Now, linearity is reaching the end of the line: designers are looking to the loop and redefining refuse as resource.

Circularity is at the core of eco-design, the production methodology in which waste is repurposed and environmental impacts such as raw-material use are reduced through reuse and recycling. But if that loop is a lasso for reining in excess, the reality — as US philosopher Ralph Waldo Emerson wrote in the industrializing 1840s — remains that "Things are in the saddle,/And ride mankind". The scale of global waste and its proportionate economic and environmental costs is gargantuan.

Some 269,000 tonnes of plastic litter the world's oceans, and vast industrial cast-offs such as manure lagoons and slag heaps blight landscapes. What lurks beneath is daunting. Landfill swallows much domestic and construction waste, where residual energy is lost and decomposition under anaerobic conditions creates a stream of problematic subwaste, from the powerful greenhouse gas methane to leachable contaminants such as benzene. The United States sends 40% of its food to landfill and discards 70–80% of the 145 million tonnes of construction and demolition debris that it generates each year — even though much of the wood, metal and minerals is recyclable. In 2012, Europe sent almost half of its 2.3 billion tonnes of waste to landfill. And that is just stuff: up to 50% of industrial energy input becomes waste heat.

Faced with this entrenched dynamic, how can closed-loop systems become the norm? One answer is to integrate them into circular economies — wheels within wheels. This model looks to extend the life of products at the 'use' stage, retaining value and designing out harmful by-products such as toxic substances, to create the perfect habitat for ecologically innovative companies.

For a model that slots so neatly into eco-thinking, the circular economy is a surprisingly venerable concept. In 1966, economist Kenneth Boulding hatched the idea of "a stable, closed-cycle, high-level technology" in his seminal paper 'The economics of the coming spaceship Earth' (see A. Rome *Nature* **527,** 443–444; 2015). Five years later, in a *Life* magazine interview, systems theorist R. Buckminster Fuller — an advocate of 'more with less' design from the 1920s — declared that pollution "is nothing but resources we're not harvesting. We allow them to disperse because we've been ignorant of their value." That year also saw the publication of *Design for the Real World* (Pantheon), an influential manifesto by Viennese educator (and ally of Fuller) Victor Papanek, who inveighed against designers creating "whole species of permanent garbage to clutter up the landscape" and called for a socially inclusive, environmentally responsible design ethic.

The 1970s saw significant practical developments. US landscape architect John T. Lyle pioneered 'regenerative design' focused on local, renewable resource use. Swiss architect Walter Stahel (see page 435) codified existing ideas and developed key new ones as principles for his Product-Life Institute in Geneva in the 1980s. More recently, German chemist Michael Braungart and US architect William McDonough (who studied under Lyle) established the product and system certification Cradle to Cradle (a coinage of Stahel's), which treats industrial flows as metabolic and waste as nutrients (C. Wise *et al. Nature* **494,** 172–175; 2013). Their book *Cradle to Cradle* (North Point) was published in 2002.

Such design revolutions are essentially longitudinal collaborations between generations, as historian of technology Walter Isaacson has revealed (J. Light *Nature* **514,** 32–33; 2014). Meanwhile, eco-design has moved on from the isolated gizmos and warranties of the 1970s, such as Germany's 'life cycle' eco-label, Blue Angel. New ventures are designing circularity in from the off, as the case studies here demonstrate. Enterra in Vancouver, Canada, recycles unsold organic food to feed fly larvae, which it then harvests as livestock feed (see 'Transform waste into protein'). AeroFarms in Newark, New Jersey, grows up to 4 million kilograms of baby ▶

▸ leafy greens a year in vertical indoor 'fields', without pesticides and using 95% less water than in field farming.

A number of grand old companies are retrofitting circularity. BAM Construct UK (of the Dutch Royal BAM Group, founded in 1869) focuses on disassembly — ensuring that the raw materials it uses are either interchangeable or easily separated, and that components can be dismantled (see 'Design for deconstruction'). UK aerospace-engine powerhouse Rolls-Royce plc has cut raw-material use, cost and emissions through its recycling programme, Revert (see 'Create consistent supply systems'), which emphasizes 'power by the hour' and remanufacturing.

Academia and governments are also waking up to circular thinking, from China (see page 440) to Europe. British sailor and circumnavigator Ellen MacArthur aims to speed the transition through her eponymous foundation in Cowes, UK, which has synthesized existing knowledge to educate on, and catalyse innovation towards, the circular economy, collaborating energetically with businesses as well as design and engineering universities. On board are Delft University of Technology in the Netherlands; the University of Bradford, UK, which established the first circular-economy master's degree in 2013; and, under a fellowship with the philanthropic US Schmidt Family Foundation in Boca Raton, Florida, a consortium of 12 universities including the Massachusetts Institute of Technology in Cambridge, Tongji University in Shanghai, China, the Indian National Institute of Design in Ahmedabad and Imperial College London.

Collectively, all this constitutes a great deal more than a gleam in Buckminster Fuller's eye. Yet if the circular economy is an ecosystem for green innovation, it is primarily an island one: wildlife corridors are few. No city, region or country has embraced the vision fully. And the urbanizing, consuming and wasting world does not stand still: the Organisation for Economic Co-operation and Development estimates that the global middle class (with all its material hankerings and 'disposable' income) will swell to 4.9 billion by 2030 (from 1.8 billion in 2009). Meanwhile, the evolving industrial worldscape — a welter of start-ups, monocultures and multinationals, most clinging to business-as-usual — contributes a dynamic unpredictability.

There are problems, too, with the circular model itself. Martin Charter, director of the Centre for Sustainable Design at the University for the Creative Arts in Farnham, UK, notes a "lack of overall clarity over the concept. Perhaps just 100 companies worldwide have adopted a true circularity mindset as a core strategy." As for the circular mantra of switching to the digital, data centres waste an average of 90% of the energy that they consume (30 billion watts, equivalent to the output of 30 nuclear power plants) and account for 17% of technology's carbon footprint. Although the circular 'business case' looks remarkable (global management consultants McKinsey and Company estimate that it could add US$2.6 trillion to the European economy by 2030), the fact that business remains central to the vision is a vulnerability. The growth economy impedes sustainability. In 2014, for instance, Chevron and a number of other big oil companies retreated from investments in renewables because of poor returns. Business competitiveness and 'disruption' can hinder the collaboration that is central to eco-design. UK design engineer Chris Wise has noted that the practice of using 'least materials' is at odds with the construction industry's prime aim of selling more materials (C. Wise *et al. Nature* **494,** 172–175; 2013). The 'rebound effect', in which designed efficiency leads to greater use or consumption, is a related conundrum.

The thirteenth-century artist Giotto reportedly proved his genius by drawing a perfect circle. The cycles of the biosphere, from water to soil, are wonders of economy. So the idea of a circle strikes a deep chord in us. But one look at any large city reveals disconnection, pollution and social inequality. Can we square the circular economy?
**Barbara Kiser**

ROLLS ROYCE

Engineers work on a Rolls-Royce BR725 engine.

## ANDREW CLIFTON
# Create consistent supply systems

*Sustainability manager — engineering and design, Rolls-Royce, Derby, UK.*

The increasing pressure placed on resources through population growth and the rising demand for energy creates a challenge for industries reliant on consistent supply of materials. Rolls-Royce — which designs, develops, makes and services integrated power systems for use in the air, on land and at sea — meets the challenge with an advanced recycling programme called Revert. This is a collaborative effort between Rolls-Royce and its material suppliers that has been implemented across 100 manufacturing facilities.

> "Almost half of a used aircraft engine can be recycled and safely used to make a new engine."

Products for aerospace applications have to withstand extreme operating conditions. Components of an aircraft's gas-turbine engines will experience temperatures ranging between −40 °C and 2,000 °C; during take-off, the loadings on the engines' front fan are equivalent to suspending 9 double-decker buses from each blade.

Such demands necessitate the use of alloys of exotic metals such as rhenium, hafnium, nickel and titanium to provide the performance, efficiency and weight savings required for today's advanced aircraft engines. Rolls-Royce uses more than 20,000 tonnes of these alloys each year, and to safeguard supply of strategic material and reduce costs it is continually working to recycle as much as possible. But recycling materials for reuse as aerospace components is not as simple as mainstream recycling, such as that of aluminium cans or scrap steel. The necessarily high quality of the material and the complexity of the alloys requires many additional safeguards if the material is to be recycled for reuse.

Enter Revert. Begun more than a decade ago, the programme is designed to help reduce costs and risks while also reducing environmental impact and safeguarding material supply. Through Revert, metal removed during the manufacture of components and from unserviceable engine parts is collected, segregated by specific alloy type, cleaned of all coatings and contaminants, and returned to the material supplier for recycling. This additional level of stewardship produces a very high-quality recyclate with the necessary chain of custody and certification for the material supplier to be able to process it back into aerospace-grade alloys.

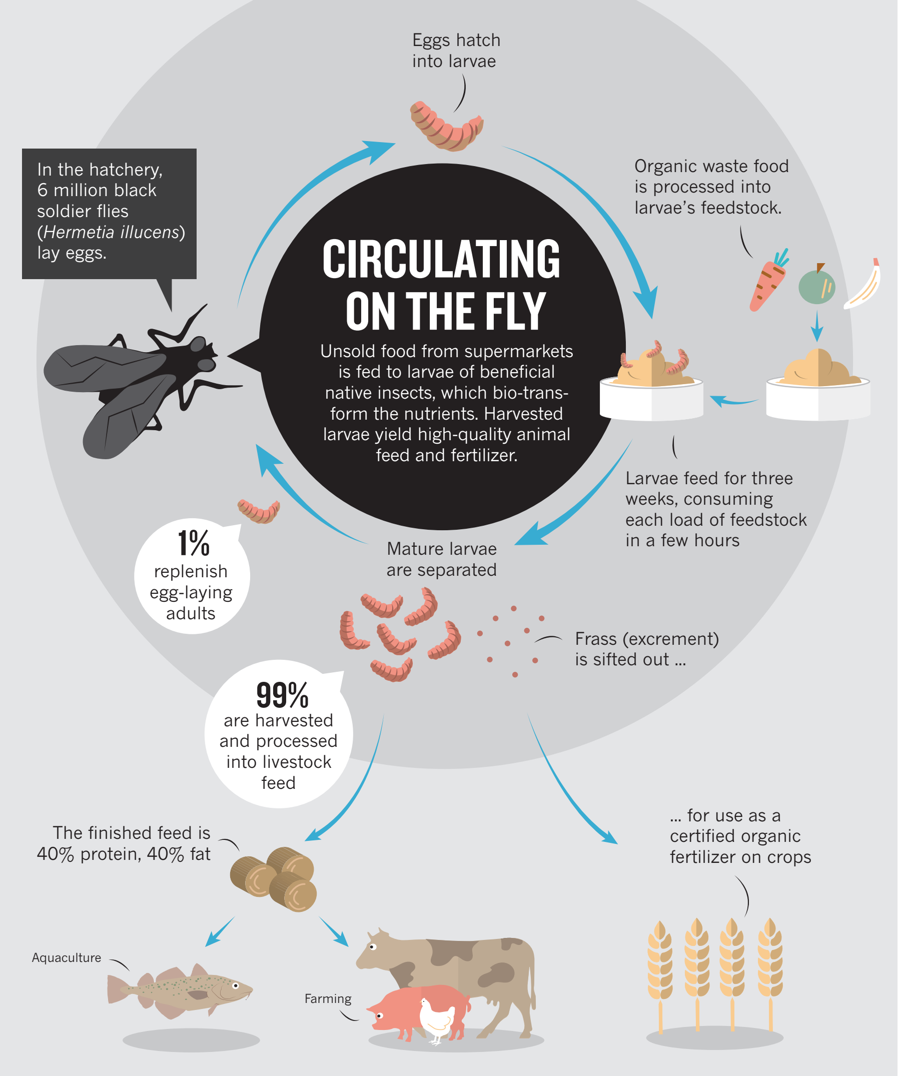
Revert is thus a triple win — providing value to the supplier, the user and the environment. The material supplier benefits from a reliable source of high-quality material to feed back into their production processes. Rolls-Royce benefits by securing long-term agreements with material suppliers that safeguard supply in exchange for the return of the Revert material. Society and the environment benefit through a reduction in environmental impact, and job creation — Revert has created some 60–70 jobs local to the Rolls-Royce site in Derby, UK, to collect and process material. The programme has cut demand for virgin material, generating energy savings of more than 300,000 megawatt-hours per year (equivalent to powering 27 million homes for a day), and a reduction in carbon dioxide emissions of 80,000 tonnes per year (equivalent to the amount emitted by the average family car circumnavigating the world 13,000 times).

As a result of Revert, much of the material used by Rolls-Royce can be reused as part of a closed-loop system. Between 90% and 100% of the titanium and nickel alloys removed during machining operations, such as milling and turning, are captured and reprocessed back into aerospace-quality material. In addition, almost half of a used aircraft engine can be recycled such that the recovered material is of high enough quality to be safely used to make a new engine. Any metallic materials that cannot be Reverted, owing to either cost or limitations in technology, are recycled as part of mainstream recycling programmes local to Rolls-Royce operations.

Revert is making a big difference. It is reducing Rolls-Royce's demand for raw materials, and it is significantly lowering costs, energy use and greenhouse-gas emissions.

## ANDREW VICKERSON
# Transform waste into protein

*Chief technology officer, Enterra Feed Corporation, Vancouver, Canada.*


CIRCULATING ON THE FLY
Unsold food from supermarkets is fed to larvae of beneficial native insects, which bio-transform the nutrients. Harvested larvae yield high-quality animal feed and fertilizer.
In the hatchery, 6 million black soldier flies (*Hermetia illucens*) lay eggs.
Eggs hatch into larvae
Organic waste food is processed into larvae's feedstock.
Larvae feed for three weeks, consuming each load of feedstock in a few hours
Mature larvae are separated
1% replenish egg-laying adults
Frass (excrement) is sifted out ...
99% are harvested and processed into livestock feed
The finished feed is 40% protein, 40% fat
... for use as a certified organic fertilizer on crops
Aquaculture
Farming

SOURCE: ENTERRA FEED CORPORATION

Fish farms and other livestock-production systems use wild forage fish such as herring and anchovies as feed ingredients in the form of fish meal and fish oil. However, as the Food and Agriculture Organization of the United Nations reports, many of these forage fish species are fully exploited, and some are depleted.

When environmentalist David Suzuki and engineer Brad Marchant met in 2007 on a rafting trip in northern Canada, Marchant wondered about alternative sources of feed. Pointing to the end of his fishing line, Suzuki asked, "Why not insects?" That idea sparked the creation of Enterra Feed Corporation in Vancouver, Canada, of which Marchant is chief executive.

Marchant, who has a track record of creating start-up companies in industrial applications of biology, saw insects as potentially solving two major global problems — a lack of sustainable feed, and wasted food. (Most of the complex nutrients in organic waste end up in landfill, compost or waste-to-energy facilities, or anaerobic digesters.) Insect larvae, he realized, could become part of a closed loop, consuming recycled food and being harvested to create a renewable source of nutrients for livestock. Feeding waste to insects allows nutrients to be recovered and used as a valuable source of protein and fat, naturally bioconverted (see 'Circulating on the fly').

PAUL HUDLOW

**The black soldier fly can turn waste food into feedstock and fertilizer.**

Unsold food — primarily fruit, vegetables, breads and grains from local grocers and food processors — arrives every day at Enterra's Vancouver facility. This feedstock is mixed in large tanks to produce a balanced diet for the large, protein-rich larvae of the black soldier fly (*Hermetia illucens*). At current capacity, the larvae at Enterra's facility can consume up to 100 tonnes of food a day. The adult fly, which has non-functional mouthparts, does not bite or even eat: it relies instead on energy stored during the larval stage to fly and reproduce.

The 6 million adult flies in the hatchery produce a constant supply of eggs. Once hatched, the larvae are fed daily for about three weeks. Each load of feed is consumed within a few hours — a fraction of the weeks or months needed to break down food in composting or waste-to-energy facilities.

Once the larvae are ready to be harvested, they are mechanically sifted to winnow out the 'frass', or manure. This is treated separately as a natural fertilizer certified for use in organic crop production. Approximately 1% of the harvested larvae are returned to the hatchery to produce more flies and eggs. The rest are processed into feed; dried, heat treated and packaged in bulk, they contain 40% protein and 40% fat, and can be shipped as is as a source of protein and oils, or turned into separate meal and oil. The larvae meal can be used in animal feed as a direct substitute for resource-intensive ingredients such as fish meal and soya-bean meal.

More than 70% of Enterra's sales have been to the United States. In January, the company received approval from the US Food and Drug Administration and the American Association of Feed Control Officials Ingredients Definition Committee to use the dried larvae as feed in salmon farming. Approval for poultry and other livestock are expected shortly. The path to approval has been slower on home territory; an application has been pending with the Canadian Food Inspection Agency since 2012. Enterra hopes to sell its products in Canada to truly close the loop on recycled food on a local scale, by producing renewable nutrients for the local feed industry, using locally sourced inputs. With US$7.5 million in capital spent and 21 full-time jobs created so far, Enterra plans to enter into partnerships globally, including in Europe, the United States, South America and Asia.

NITESH MAGDANI

## Design for deconstruction

*Director of Sustainability, BAM Construct UK, Hemel Hempstead.*

With sustainable construction company BAM Construct UK, I help to develop buildings that are fit for purpose and perform as intended for their whole lifetime. My focus is passive design, renewable and modular materials and buildings with low energy demand. Major influences have been the Cradle to Cradle approach and architects who explore biomimicry, such as Antoni Gaudí and Santiago Calatrava. The circular-economy model has provided a larger organizing idea with which to synthesize these strands, because it is predicated on using materials that can retain asset value for longer and can eventually be taken back to their biological or technical cycles — reused, repurposed or remanufactured — to reduce waste and unlock new economic opportunities.

BAM's first 'circular' pilot project is the town hall in Brummen, the Netherlands. The client had outgrown its existing building and needed a larger space for at least another 20 years. With Rau Architects in Amsterdam and its sister company Turntoo, we offered a 'building as material bank' to maximize value for the municipality (given that it may wish to move its offices in time). Our competition-winning offer took into account the full costs of the building over its 20-year occupancy, and provided greater price certainty than conventional approaches. Key to this was that after 20 years, components of the building (such as structural timber and metals) could be returned, under contract, to suppliers, unlocking a minimum 20% of their residual value. This 'closed loop' approach reduces manufacturers' reliance on virgin materials and diminishes price volatility.

The town hall in Brummen, the Netherlands, was built to be disassembled and recycled.

BAM BOUW EN TECHNIEK

Technical elements designed for disassembly include the overall shell, cladding, internal partitions and cooling. The design avoids coatings and resins wherever possible to make parts interchangeable and allow separation of valuable raw materials. Components have to retain value over time, so we bring partners such as electronics suppliers and manufacturers Philips and 8Point3 to the table. Many of our projects also incorporate prefabricated elements, so design proceeds through a standardized procurement process to reduce production costs, as well as increasing residual values for key components to more than 50%.

> *"Transparency through the supply chain is essential, so our work has to be highly collaborative."*

Beyond materials, we look at systems and processes such as the cost of dismantling, logistics, and storage of components, how it is done, and by whom. We decide who takes responsibility and ownership of the materials during and after the use phase. Transparency through the supply chain is essential, so by its very nature, our work has to be highly collaborative. Sander Holm, a key sustainability leader at our Dutch sister construction and engineering company BAM Bouw en Techniek, notes, "manufacturers and suppliers must sit at the table together as soon as possible. This kind of co-creation delivers more innovation, and also a higher residual value."

We are currently starting a project with The Great Recovery, a sustainability network launched in 2012 by the RSA (formerly the Royal Society of Arts) in London, to encourage designers, manufacturers and recyclers to co-create solutions for material reuse. We are using 'teardown' methodology, in which production systems are scrutinized to tease out problems and opportunities for 'designing up' to circularity. BAM will focus on processes key to the circular economy, including building-information modelling, a digital shared-knowledge resource used to make decisions about a building's life cycle, such as resource productivity.

At the moment, there is no guarantee that buildings or products designed with natural materials or for deconstruction will be reused. And there is little information on existing building stocks and their potential for sustainable renovation. This must change. One of BAM's challenges is the need to educate the value chain — encouraging our industry towards procurement with an eye to the longer term, and switching to 'performance or take-back' contracts, which keep the responsibility for maintenance, durability and replacement of parts with suppliers. Through the UK Supply Chain Sustainability School, BAM has hosted the first of a series of workshops to work through some of the barriers to circular-economic models. Building on our status as leader in the Dutch Benchmark Circular Business Practices 2015, we are gradually moving from focusing on waste reduction in the construction process to reducing waste over a building's life cycle. ■

# Correspondence

## Unlimited by-catch limits recovery

We argue that government decisions to increase the social and economic benefits of fisheries will be ineffective without improvements in data-reporting practices and in regulations for targeted by-catches (see J. Casey *et al. Nature* **530,** 160; 2016).

Take the swordfish (*Xiphias gladius*), a target species in the Atlantic longline fishery with a strictly regulated low annual total allowable catch. The unregulated 'by-catch' consists mainly of shortfin mako (*Isurus oxyrinchus*) and blue sharks (*Prionace glauca*), whose fins and meat are commercially valuable. Similarities between the gutted carcasses of swordfish and mako, without heads or fins, mean that there is potential for illegal overfishing of a regulated species that can then be logged as an unregulated species on landing.

Catches of shortfin mako typically comprise 3–13% of blue-shark catches in the same longline or gill-net fishery (J. D. Stevens in *Sharks of the Open Ocean* Ch. 7, 90; Blackwell, 2008). Yet the 2008 mako landings of a European fleet were, on average, six times those of blue shark (unpublished data; available from D.W.S. and N.Q.). This implies that the excess 'mako' could have been a regulated species such as swordfish.

The scale of the problem may already have affected stock rebuilding. More stringent surveillance of by-catch species by national regulatory authorities is essential for spotting such irregularities.

**David W. Sims** *Marine Biological Association, Plymouth, UK.*
**Nuno Queiroz** *CIBIO/InBIO — University of Porto, Portugal.*
*dws@mba.ac.uk*

## Keep allowable fish catches sustainable

Setting total allowable catches (TACs) for European fish stocks above those advised by the International Council for the Exploration of the Sea does not necessarily mean that stocks are being overfished (see *Nature* **528,** 435; 2015).

The status of assessed stocks is in fact improving in some European Union regions that are managed through TACs (north-east Atlantic, North Sea, Baltic Sea), although overfishing is still evident in the Mediterranean and Black seas (see go.nature.com/ojr1ue). These trends indicate that we know how to achieve sustainable fisheries.

Sustainable yields and the political will to achieve them will ultimately determine the socio-economic viability of the fishing industry. Although short-term socio-economic factors often drive EU fisheries policy (see J. Casey *et al. Nature* **530,** 160; 2016), there are alternatives. In the United States, for example, these factors may be used only to reduce quotas, not increase them (Magnuson–Stevens Act, 2007).

**Griffin Carpenter** *New Economics Foundation, London, UK.*
**Sebastian Villasante** *University of Santiago de Compostela, A Coruña, Spain.*
**Bethan C. O'Leary** *University of York, UK.*
*griffin.carpenter@neweconomics.org*

## SDG indicators need crowdsourcing

The Inter-agency and Expert Group on Sustainable Development Goal Indicators (IAEG-SDGs) meets at the end of this month to establish, among other things, the development of global reporting mechanisms (http://unstats.un.org/sdgs). We suggest that data that have been crowdsourced by civil-society ventures should be incorporated into the international process of monitoring the SDGs.

To address potential issues of data quality, initiatives such as the Open Seventeen Challenge are necessary to train organizers of crowdsourcing projects (see http://openseventeen.org). This initiative draws on advice from leaders in advocacy, governance and crowdsourcing tools. It is run by Citizen Cyberlab, a partnership between the University of Geneva, the United Nations Institute for Training and Research (UNITAR) and Europe's particle-physics lab CERN.

For crowdsourcing to achieve its full potential, governments will need to support projects that promote public participation in measuring progress towards the SDGs. National statistics offices must develop best practices for integrating crowdsourced data.

As a first step, we encourage the IAEG-SDGs to emphasize crowdsourcing as a legitimate and valuable contribution to tracking the goals.

**Yves Flückiger** *University of Geneva, Switzerland.*
**Nikhil Seth** *UNITAR, Geneva, Switzerland.*
*citizencyberlab@unige.ch*

## Protect Czech park from development

The Czech Republic's parliament will vote this month on a bill that sets new rules for national parks. In eastern Europe's push for economic development, biodiversity is again under threat (see also P. Chylarecki and N. Selva *Nature* **530,** 419; 2016; P. Michalak *Nature* **530,** 419; 2016).

Šumava National Park is a unique complex of peat bogs, wetlands and primeval forests in southern Bohemia. It is a refuge for many endangered species, including the remaining few viable populations in central Europe of capercaillies (*Tetrao urogallus*) and freshwater pearl mussels (*Margaritifera margaritifera*).

The bill, a reasonable compromise for conservation, is under attack from regional politicians. They object to the proposed transparency in setting and implementing conservation rules, and seek to restrict protected core zones to 23% of the national-park area. They also want to open up large areas for logging, tourism and privatization.

Their arguments are similar to those used to justify logging in Białowieża Forest on the Poland–Belarus border (see *Nature* **530,** 393; 2016). However, if Šumava's rare species are to survive, core zones need to cover about 50% of the park (I. Dickie *et al. Eur. J. Environ. Sci.* **4,** 5–29; 2014).

**Pavel Kindlmann, Zdenka Křenová** *Global Change Research Institute, Brno; and Charles University, Prague, Czech Republic.*
*pavel.kindlmann@centrum.cz*

## Matchmaker aims to cut journal shopping

Scientific publishing is getting slower, in part because authors often choose to submit a paper to successive journals until one agrees to publish (*Nature* **530,** 148–151; 2016). As an unpaid academic member of the editorial board of Axios Review, a private, third-party review service, I wish to point out that a simple solution already exists for the time-wasting problem of 'journal shopping'.

Journals do not permit simultaneous submission to other journals, but organizations such as Axios Review — others include Peerage of Science and Rubriq — use peer review to assess the 'fit' of a paper to multiple journals and then pass it to the most suitable outlet (go.nature.com/dhg6hn).

This approach is based on the principle of parallelization — a solution to delays caused by serial events. Programmers use parallel processors to enable faster computation, for example, and parallel DNA sequencing has increased the output of genetic data. Early evidence suggests that parallelization also significantly shortens the review process: 85% of papers peer-reviewed through Axios Review, for instance, get accepted by the first journal.

**Robert H. S. Kraus** *University of Konstanz, Germany.*
*robert.kraus@uni-konstanz.de*

ALZHEIMER'S DISEASE

# Lost memories found

**Enhancing synaptic connections between neurons in the brain's hippocampus that are normally activated during memory formation rescues memory deficits in a mouse model of early Alzheimer's disease.** SEE LETTER P.508

PRERANA SHRESTHA & ERIC KLANN

Some 44 million people worldwide are affected by the neurodegenerative disorder Alzheimer's disease or related dementia[1]. Episodic memory decline in the early stages of disease often precedes biological hallmarks such as the abnormal accumulation of amyloid-β and tau proteins. It is also better correlated than these traits with defects that occur in the synaptic connections between neurons in patients' brains[2]. However, whether this memory impairment reflects a failure to encode new memories or an inability to retrieve stored ones is not known. In a fascinating paper on page 508 of this issue, Roy *et al.*[3] shed light on the matter, rescuing synaptic and memory deficits in a mouse model of early Alzheimer's disease by reactivating an ensemble of neurons that had been previously activated by experience.

Memory is a biological function that allows animals to encode, retain and retrieve information. The brain's hippocampus is a key player in these processes. Hippocampal damage can be assessed in mice using a behavioural model called contextual fear conditioning, in which animals learn to show defensive behaviour in a certain environmental context after it becomes associated with an aversive stimulus — for example, learning to freeze when placed in a box in which they have previously received an electric shock to the foot. Roy and colleagues found that mouse models of the early stages of Alzheimer's disease can encode memories of an aversive experience, but that their recall of the experience when placed in the same environment a day later is severely impaired (Fig. 1a).

Neuronal ensembles that undergo enduring biochemical changes during an experience, and that are reactivated during recall of that experience, are referred to as engram cells[4]. Using mouse models of early-stage Alzheimer's disease, Roy *et al.* modified a previously described protocol[5] so that engram cells in the dentate-gyrus subregion of the hippocampus that are activated by an aversive experience are tagged with a gene that encodes a light-sensitive ion channel. The authors then used blue light to selectively open the channel, activating the engram (the population of engram cells in the dentate gyrus) and so manipulating the memory trace at a later time.

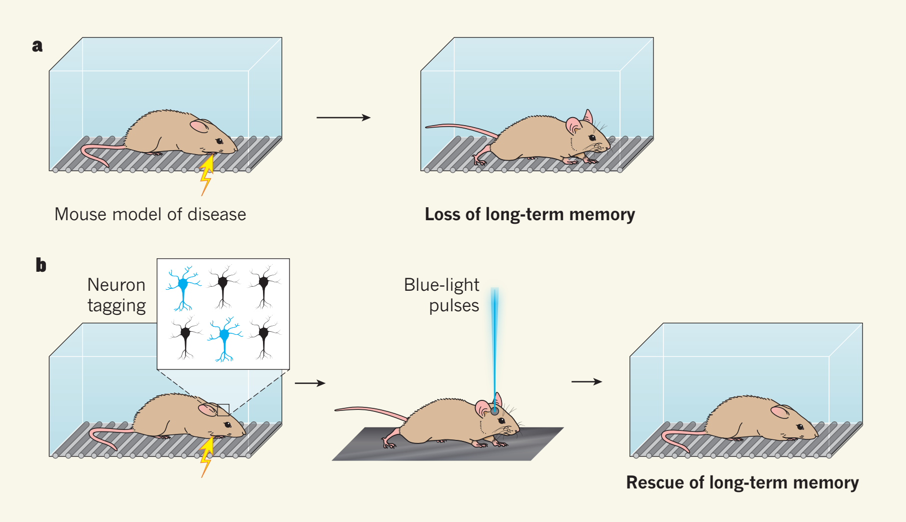


**Figure 1 | Light-induced rescue of episodic memory. a**, During contextual fear conditioning, a mouse is placed in a box where it undergoes an aversive experience, such as a footshock (yellow arrow), which causes it to freeze. Healthy mice recall this experience and freeze when placed back in the same box (not shown), but mice that model Alzheimer's disease do not, indicating that they have long-term memory defects. **b**, Roy *et al.*[3] genetically tagged neurons that were activated during the aversive experience with a light-sensitive ion-channel protein (tagged neurons indicated in blue). They then removed the mice from the fear-conditioning environment and used pulses of blue light to repeatedly reactivate these neurons. In doing so, the authors rescued long-term memory in the model mice.

Light-induced reactivation of the tagged engram cells led to recall of the aversive memory. However, memory rescue did not extend beyond the duration of the light treatment, so it could not reinstate memory recall when, a day later, the animals were returned to the context in which they experienced the footshock (known as a natural cue). This finding suggested that the standard optical reactivation regimen did not strengthen the engram enough for it to have a lasting effect. Indeed, it is known that associative memory traces need to be consolidated — strengthened and stabilized — to facilitate future retrieval in response to natural cues[6].

Consolidation of contextual fear conditioning involves long-lasting strengthening of the synaptic inputs to neurons in the dentate gyrus from neurons in another brain region, the entorhinal cortex[7]. Such strengthening is known as long-term potentiation (LTP). Roy *et al.* reasoned that the memory-retrieval deficit in the model mice came from an inability to strengthen these 'perforant path' synapses. To test this theory, they repeatedly activated the engram with serial high-frequency light pulses[8], optically inducing LTP in the perforant-path inputs.

With this clever strategy, the authors rescued long-term memory deficits in the mice (Fig. 1b). A reduction in the density of tiny neuronal structures called dendritic spines, which receive synaptic inputs, is associated with memory loss in Alzheimer's disease, and optically induced LTP returned spine density to normal. Moreover, this strategy rescued deficits in two other hippocampus-dependent memory tasks: active place avoidance, in which an animal learns to avoid the specific place where it previously received a footshock, and novel object recognition, which tests recognition memory. This indicates that the engram-based technique can be generalized to improve recall for various memory types.

Next, Roy *et al.* found that optically inducing LTP in perforant-path inputs to a broader range of dentate-gyrus neurons did not rescue long-term memory. This finding is intriguing, because it suggests that simultaneous reactivation of multiple neuronal ensembles in the dentate gyrus cancels out the effects of reactivating a specific engram. Consequently, treatments such as electrical stimulation of deep brain regions, which are used to treat human neurological disorders but cannot discriminate between engram and non-engram cells, may not improve memory in patients.

Notably, a previous study[9] showed that electrical stimulation of the perforant path increases levels of amyloid-β in the interstitial fluid around hippocampal cells. Further work is needed to determine whether Roy and colleagues' engram intervention increases amyloid-β levels, and whether the strategy can ameliorate memory impairments in late-stage Alzheimer's disease if combined with techniques[10] to reduce amyloid-β levels and aggregation of tau.

To both tag and manipulate engrams in mice, Roy *et al.* introduced genetic constructs in two viruses — a strategy that comes with caveats. One construct contained a short, 1-kilobase promoter region, which drives gene expression in active neurons. In its natural state in the genome, the promoter drives *c-Fos* expression, but in the viral construct it promotes expression of an 'activator' gene that, in turn, drives expression of a second construct that encodes the ion channel. However, this promoter naturally acts in concert with enhancer elements that span the 50 kilobases of DNA surrounding it[11]. Excluding these gene-regulatory elements from the viral construct results in an incomplete engram, because some neurons that are activated by the aversive experience less strongly than others will not be tagged. The engram could be labelled with greater specificity by incorporating the activator into the genomic position of *c-Fos*, such that all the gene-regulatory elements can act in concert to tag neuronal ensembles in response to aversive experience.

In addition, regardless of promoter expression, the construct containing the ion channel can be activated only when an antibiotic called doxycycline is removed from the animals' diet. Roy and colleagues tagged engram cells for 24 hours from the start of contextual fear conditioning. This design lacks precision, so some nonspecific neurons are probably included in the tagged ensemble. Engram labelling could be optimized by decreasing the time for which doxycycline is removed from the diet, or by using an alternative engram-tagging strategy that allows a shorter time window for labelling[12].

Nonetheless, the potential to rescue long-term memory in dementia is exciting. In the future, Roy and colleagues' findings might help to guide engram-based strategies that rescue memory deficits in patients with early-stage Alzheimer's disease. ■

**Prerana Shrestha** *and* **Eric Klann** *are in the Center for Neural Science, New York University, New York, New York 10003, USA.*
*e-mail: ps755@nyu.edu; ek65@nyu.edu*

1. Van Cauwenberghe, C., Van Broeckhoven, C. & Sleegers, K. *Genet. Med.* http://dx.doi.org/10.1038/gim.2015.117 (2015).
2. Terry, R. D. *et al. Ann. Neurol.* **30,** 572–580 (1991).
3. Roy, D. S. *et al. Nature* **531,** 508–512 (2016).
4. Tonegawa, S., Liu, X., Ramirez, S. & Redondo, R. *Neuron* **87,** 918–931 (2015).
5. Reijmers, L. G., Perkins, B. L., Matsuo, N. & Mayford, M. *Science* **317,** 1230–1233 (2007).
6. Abel, T. & Lattal, K. M. *Curr. Opin. Neurobiol.* **11,** 180–187 (2001).
7. Nguyen, P. V. & Kandel, E. R. *J. Neurosci.* **16,** 3189–3198 (1996).
8. Nabavi, S. *et al. Nature* **511,** 348–352 (2014).
9. Cirrito, J. R. *et al. Neuron* **48,** 913–922 (2005).
10. Leinenga, G. & Götz, J. *Sci. Transl. Med.* **7,** 278ra33 (2015).
11. Joo, J.-Y., Schaukowitch, K., Farbiak, L., Kilaru, G. & Kim T.-K. *Nature Neurosci.* **19,** 75–83 (2016).
12. Cazzulino, A. S., Martinez, R., Tomm, N.K. & Denny, C. A. *Hippocampus* http://dx.doi.org/10.1002/hipo.22556 (2015).

**This article was published online on 16 March 2016.**

ANIMAL MIGRATION

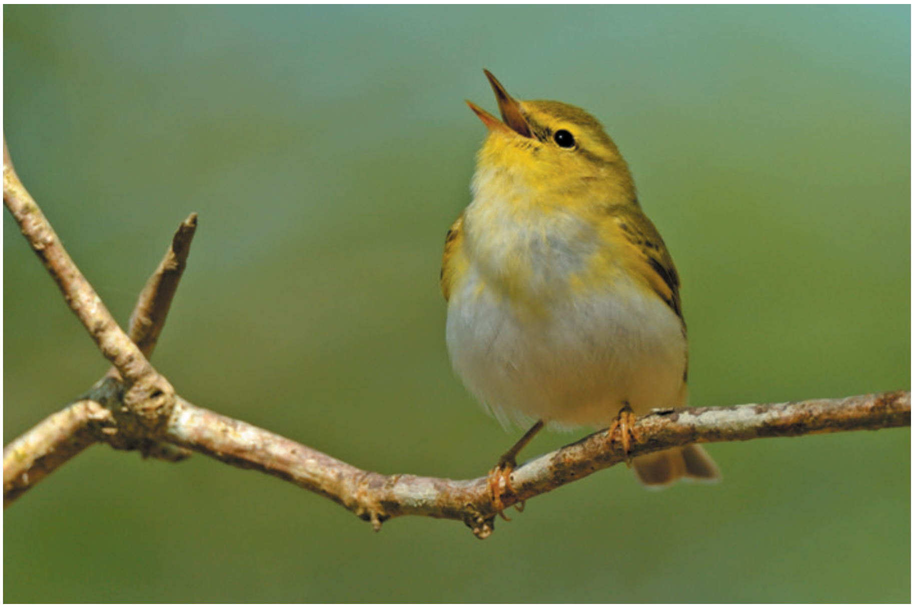
# Dispersion explains declines

**Migratory birds are declining globally. A broad study of European migratory birds finds that species that disperse widely during the non-breeding season are less likely to be in decline than are species with more restricted dispersion.**

RICHARD A. FULLER

Migratory birds undertake some of the most extraordinary journeys of any animal, but many of these birds are in catastrophic decline[1]. The very mobility of these species makes it extremely difficult to diagnose causes of the declines, and painstaking ecological studies are needed to unpick them on a case-by-case basis[1]. Writing in *Ecology Letters*, Gilroy *et al.*[2] present data hinting at a much-needed general explanation for why some migratory species are more vulnerable than others. In an analysis of 340 migratory bird species, they show that species that disperse widely during the non-breeding season, relative to their breeding distribution, are much less likely to be declining than are species that have relatively more-restricted distributions outside the breeding season.

The distances travelled by some migratory birds are astounding. The blackpoll warbler (*Setophaga striata*), a forest songbird weighing only 12 grams, flies more than 2,500 kilometres



ANDY ROUSE/2020VISION/NATUREPL.COM

**Figure 1 | Wood warbler (*Phylloscopus sibilatrix*).** Although this declining bird species has extensive breeding grounds across Europe, it spends the non-breeding season in a relatively small area in west and central Africa. Gilroy *et al.*[2] find that such low migratory dispersion is associated with population decline.

## 50 Years Ago

The American Institute of Physics is uncommonly well informed about the jobs its members do, and about the salaries they are paid. The latest batch of figures, made public in *Physics Today* for January, will as usual comfort those struggling for a Ph.D. with the knowledge that their efforts (if successful) are likely to add something like 5,000 dollars to an annual salary in industrial research and development. In 1964 the median starting salary for Ph.D. physicists in industry appears to have been 12,600 dollars, compared with 8,800 dollars for those starting with a master's degree and 7,500 dollars straight after graduation. The initial value of a Ph.D. seems to be equivalent to ten years of plodding up the promotion ladder … By contrast, academic life offers a lower starting salary but faster promotion … The profession of physics has something in common with professional football, where a man must reckon that his earning power will disappear altogether at forty.
**From *Nature* 26 March 1966**

## 100 Years Ago

The *Museums Journal* for March very properly reprints the recent discussion in the House of Lords on the closing of museums … For we have in this the measure of the value our rulers set upon the scientific work of the country. We talk much of the education of the "masses," but it is now abundantly evident that the "educated" have still much to learn. Many of the speakers during the debate seemed to be under the impression that the mental equipment attained at Eton suffices to meet all the demands of later life. Though some of the speakers were actually trustees of the British Museum, yet they displayed neither knowledge of the nature of the work of that institution, nor of museums in general.
**From *Nature* 23 March 1916**

non-stop over open ocean in its make-or-break migration from the boreal forests of the Northern Hemisphere to northern South America[3]. The bar-tailed godwit (*Limosa lapponica*) flies 12,000 km non-stop over the Pacific Ocean from Alaska to New Zealand[4], and the Arctic tern (*Sterna paradisaea*) covers the distance to the Moon and back three times during its lifetime[5].

But these remarkable journeys depend on the availability of suitable destinations. The slender-billed curlew (*Numenius tenuirostris*), which may now be extinct, migrated from breeding grounds in Siberia to tiny areas in southern Europe and North Africa, where suitable wintering habitat has rapidly declined through the conversion of wetlands to farmland[6]. Its case is potentially the first extinction of a European bird since the demise[6] of the great auk (*Pinguinus impennis*) in the mid-nineteenth century.

There are more than 1,200 migratory bird species in the world, and many may wane to rarity or extinction before we have worked out why they are in trouble. Despite this scientific uncertainty, immediate conservation action is crucial[7]. But little progress has been made in identifying general explanations for the enormous declines in migratory animals, which severely hampers effective conservation planning.

Analysing a database of Europe-wide population trends for 340 bird species, Gilroy *et al.* tested several hypotheses for why some 36% of Europe's migratory bird species are in decline. About 40% of the species studied have non-breeding ranges that are larger than their breeding ranges, which the authors term a high migratory dispersion. The researchers found that these species were less likely to be in population decline than were others, when factors such as the effects of migration distance, habitat use and climatic niche width were controlled for.

For example, the wood warbler (*Phylloscopus sibilatrix*; Fig. 1) breeds across a large swathe of Europe, from Britain to Ukraine, but spends the non-breeding season in a relatively small area in west and central Africa. This species is in rapid decline. By contrast, the Eurasian reed warbler (*Acrocephalus scirpaceus*), which has a similar breeding distribution to the wood warbler but occurs across the whole of sub-Saharan Africa in the non-breeding season, has a stable or increasing population. This kind of difference in migration strategy is a surprisingly good predictor of population declines, and is consistent with increasing concern that desertification, habitat loss and degradation in the wintering grounds of Europe's migratory birds are driving a new wave of population collapse[7].

Some species do not exhibit the full to-and-fro migration seen in the warblers that decamp wholesale to Africa, but instead are present year-round in some areas of their breeding distribution. Gilroy *et al.* found that these partial migrants were notably less likely to be declining than were full migrants, suggesting a clear advantage to this strategy.

Gilroy and colleagues' wide-ranging study also pinpointed several other influences on migratory-bird populations. They found that declines were especially pronounced among habitat specialists (particularly, farmland species), small-bodied species and, intriguingly, those that have not advanced their annual arrival date in Europe between 1960 and 2006 to start breeding in response to the earlier onsets of spring. This latter effect could result from climate-adaptable species being more resilient to decline[8], although it is also possible that species that are not declining for other reasons have greater population variability in arrival date, on which selection can operate.

Studies of other geographic regions and taxa are needed to establish the generality of these results, but for now the findings are a major step forward in predicting the possible need for conservation action among the world's migratory bird species, simply on the basis of maps of their seasonal distributions. New technology, such as the network of 300 automated telemetry towers across North America as part of the Motus project[9], is transforming our knowledge of migratory routes, and innovative analysis of threats across the annual cycle of migratory species[10] is opening up ways of planning effective conservation action.

But more data alone will not save migratory species. Ambitious global and regional conservation agreements and initiatives, such as the Convention on the Conservation of Migratory Species of Wild Animals, the African-Eurasian Migratory Landbirds Action Plan and the East Asian-Australasian Flyway Partnership, are beginning to bear fruit and achieve joined-up conservation. Because of the extreme reliance of many migrants on small areas at some point in their migratory cycle, smartly targeted conservation action may be effective in reversing population declines. ■

**Richard A. Fuller** *is at the School of Biological Sciences, University of Queensland, Brisbane, Australia.*
*e-mail: r.fuller@uq.edu.au*

1. Wilcove, D. S. & Wikelski, M. *PLoS Biol.* **6,** e188 (2008).
2. Gilroy J. J., Gill, J. A., Butchart, S. H. M., Jones, V. R. & Franco, A. M. A. *Ecol. Lett.* **19,** 308–317 (2016).
3. DeLuca, W. V. *et al. Biol. Lett.* **11,** 20141045 (2015).
4. Gill, R. E. *et al. Proc. R. Soc. B* **276,** 447–457 (2009).
5. Egevang, C. *et al. Proc. Natl Acad. Sci. USA* **107,** 2078–2081 (2010).
6. Clare, H. *Orison for a Curlew* (Little Toller, 2015).
7. Kirby, J. S. *et al. Bird Conserv. Int.* **18,** S49–S73 (2008).
8. Møller, A. P., Rubolini, D. & Lehikoinen, E. *Proc. Natl Acad. Sci. USA* **105,** 16195–16200 (2008).
9. http://motus-wts.org
10. Rushing, C. S., Ryder, T. B. & Marra, P. P. *Proc. R. Soc. B* **283,** 20152846 (2016).

ORGANIC CHEMISTRY

# No double bond left behind

**Alkenyl halides are some of the most useful building blocks for synthesizing small organic molecules. A catalyst has now allowed their direct preparation from widely available alkenes using the cross-metathesis reaction.** SEE ARTICLE P.459

DAVID SARLAH

The development of synthetic methods for chemical transformations is an ongoing challenge. Advances in this field have far-reaching effects on everyday life, because new transformations can streamline the preparation of pharmaceuticals, agrochemicals and materials. On page 459 of this issue, Hoveyda and colleagues[1] describe just such an advance: a reaction that allows direct access to *cis*-substituted vinyl fluorides, chlorides and bromides, members of the alkenyl halide family of compounds that are of great value in organic synthesis.

Organic molecules that contain halogen atoms, such as alkenyl halides — characterized by the C=C–X group, where X is a halogen atom such as fluorine, chlorine or bromine — have a central role in the rapid construction of organic molecules in general. This is because many of the most useful synthetic reactions are catalysed by transition metals, which insert into carbon–halogen bonds and thus enable the carbon atom to form bonds with other atoms[2]. Certain alkenyl fluorides, chlorides and bromides are also essential structural motifs for the biological activities of drugs and naturally occurring compounds[3,4].

However, some alkenyl halides are not easy to prepare. Olefination reactions are known that form carbon–carbon double bonds (C=C bonds) with halogens attached, but these frequently produce substantial amounts of side products[5]. Not all alkenyl halides can be obtained in this way, and even when this approach is successful, the products are often isolated as an undesirable mixture of *cis*- and *trans*-isomers (in *cis*-isomers, the groups are on the same side of the double bond and point in the same direction, but in *trans*-isomers they are on different sides of the double bond and point in opposite directions). Alkenyl halides can also be made by reacting a positively charged halogen species with alkenes (hydrocarbons that incorporate double bonds) containing boron[6], silicon[7] or tin atoms[8], but the preparation of these alkenes is cumbersome.

An alternative synthetic strategy that could convert readily available alkenes to their corresponding alkenyl halides is cross-metathesis, a variant of the olefin metathesis transformation. Olefin metathesis is a fundamental organic reaction that redistributes C=C bonds, providing a simple route to a wide range of compounds; cross-metathesis involves the redistribution of double bonds between two different alkene molecules (Fig. 1a). Advances in olefin metathesis have greatly expanded the chemist's toolbox of carbon–carbon bond-forming reactions[9] because of the versatility of this process, and because of the abundance of alkenes.

But, so far, efforts to use cross-metathesis to convert alkenes to alkenyl halides have met with little success. Such reactions require different alkenyl halides to be used as starting materials, but these compounds tend to deactivate metathesis catalysts by forming several species that drastically slow down, or even stop, the catalytic cycle. For example, attempts to use standard ruthenium-based catalysts in such reactions result in the formation of species (carbenes and carbides) that are catalytically inactive[10].

Hoveyda and colleagues have circumvented this problem by exploring the use of catalysts based on tungsten and molybdenum in high oxidation states. The authors report that a molybdenum-based catalyst has outstanding activity in cross-metathesis reactions, and delivers high yields of a range of vinyl halides (Fig. 1b).

Most alkenyl halides are toxic gases at ambient conditions, so the researchers instead used commercially available liquid 1,2-dihaloethene reagents — alkenyl halides in which a halogen is attached at both ends of a C=C bond — as reactants to convert alkenes to vinyl chlorides and bromides (Fig. 1b). This protocol is a great advance because it makes the process operationally convenient and safe. The authors also used *Z*-bromo-fluoroethene to make vinyl fluorides, a reaction that could, in principle, also produce vinyl bromides as side products. Impressively, the reactions generated vinyl fluorides selectively, especially for molecules in which the C=C bond is close to a bulky group. The authors propose that steric effects (associated with the spatial crowding of chemical groups) and electronic factors account for this remarkable selectivity.

Furthermore, all the vinyl halides made in the authors' reactions were prepared with high *cis*-selectivity, a characteristic feature of this type of molybdenum catalyst that Hoveyda's group established previously[11]. The highest *cis*-selectivities were obtained in reactions of alkenes that contain bulky groups, but alkenes with linear chains also yielded synthetically useful *cis*:*trans* ratios of products. Finally, the authors showcased their chemistry by using it to convert a cyclic alkene — a ring of eight carbons that includes one C=C bond — into a linear molecule with a vinyl bromide at either end (see Fig 3. of the paper[1]). This compound has been used[12] in the synthesis of

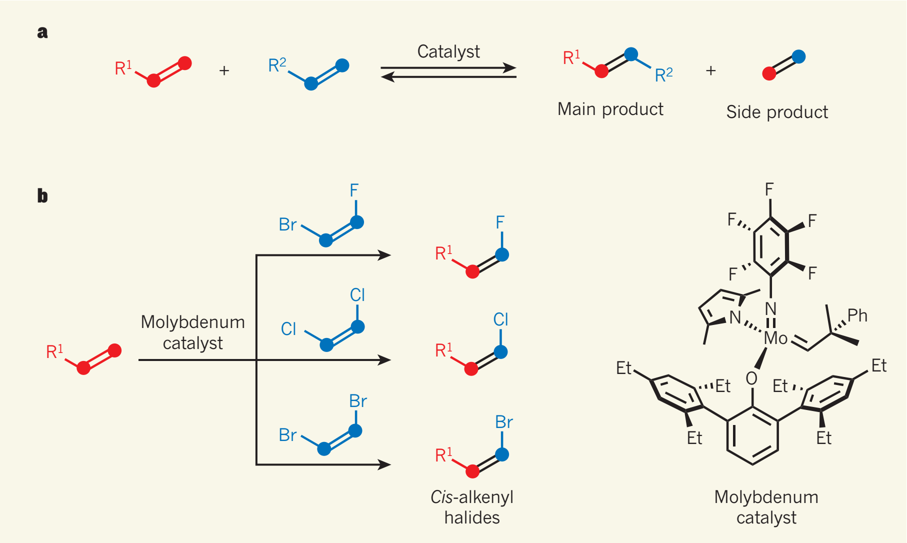


**Figure 1 | Cross-metathesis to make *cis*-alkenyl halides.** **a**, In cross-metathesis, carbon–carbon double bonds (C=C bonds) are redistributed between two molecules. Dots indicate carbon atoms; $R^1$ and $R^2$ represent any chemical group. **b**, Hoveyda and colleagues[1] report that a molybdenum catalyst allows alkenes (red) to react with 1,2-dihaloethene reagents (blue) to produce *cis*-alkenyl halides as the main products; for simplicity, side products are not shown. Et, ethyl group; Ph, phenyl group.

tetrahydrosiphonodiol, a natural product that has antitumour activity.

Hoveyda and colleagues' transformation offers a powerful strategy for preparing *cis*-alkenyl halides, especially given that alkenes are abundant motifs in fine chemicals such as pharmaceuticals and agrochemicals. In addition, the tolerance of the reaction to a broad range of chemical groups could inspire synthetic chemists to apply this reaction in more complex molecular settings: for example, in intermediates prepared during the late stages of synthetic routes to drug candidates, or in natural-product synthesis.

It should be noted that the air-sensitive catalyst is not commercially available at present, nor easy to synthesize. However, the authors say that paraffin capsules containing the catalyst will become commercially available and could be used outside the nitrogen-filled boxes commonly required for handling air-sensitive reagents, which would greatly simplify the catalyst's use. Further mechanistic insights will be needed to explain why the mono-halomethylidene species formed as catalytic intermediates in the reactions are so much more reactive than those formed in previous attempts at these reactions. However, there is little doubt that the influence of this transformation will further increase the already widespread application of cross-metathesis in chemical synthesis. ■

**David Sarlah** *is at the Roger Adams Laboratory, Department of Chemistry, University of Illinois, Urbana, Illinois 61801, USA.*
*e-mail: sarlah@illinois.edu*

1. Koh, M. J., Nguyen, T. T., Zhang, H., Schrock, R. R. & Hoveyda, A. H. *Nature* **531,** 459–465 (2016).
2. Jana, R., Pathak, T. P. & Sigman, M. S. *Chem. Rev.* **111,** 1417–1492 (2011).
3. Silverman, R. B., Bichler, K. A. & Leon, A. J. *J. Am. Chem. Soc.* **118,** 1253–1261 (1996).
4. Renner, M. K., Jensen, P. R. & Fenical, W. *J. Org. Chem.* **63,** 8346–8354 (1998).
5. Hodgson, D. M. & Arif, T. *J. Am. Chem. Soc.* **130,** 16500–16501 (2008).
6. Brown, H. C., Hamaoka, T. & Ravindran, N. *J. Am. Chem. Soc.* **95,** 6456–6457 (1973).
7. Pawluć, P., Hreczycho, G., Szudkowska, J., Kubicki, M. & Marciniec, B. *Org. Lett.* **11,** 3390–3393 (2009).
8. Nicolaou, K. C., Veale, C. A., Webber, S. E. & Katerinopoulos, H. *J. Am. Chem. Soc.* **107,** 7515–7518 (1985).
9. Nicolaou, K. C., Bulger, P. G. & Sarlah, D. *Angew. Chem. Int. Edn* **44,** 4490–4527 (2005).
10. Macnaughtan, M. L., Johnson, M. J. A. & Kampf, J. W. *J. Am. Chem. Soc.* **129,** 7708–7709 (2007).
11. Meek, S. J., O'Brien, R. V., Llaveria, J., Schrock, R. R. & Hoveyda, A. H. *Nature* **471,** 461–466 (2011).
12. López, S., Fernández-Trillo, F., Midón, P., Castedo, L. & Saá, C. *J. Org. Chem.* **70,** 6346–6352 (2005).

MICROBIAL OCEANOGRAPHY

# Viral strategies at sea

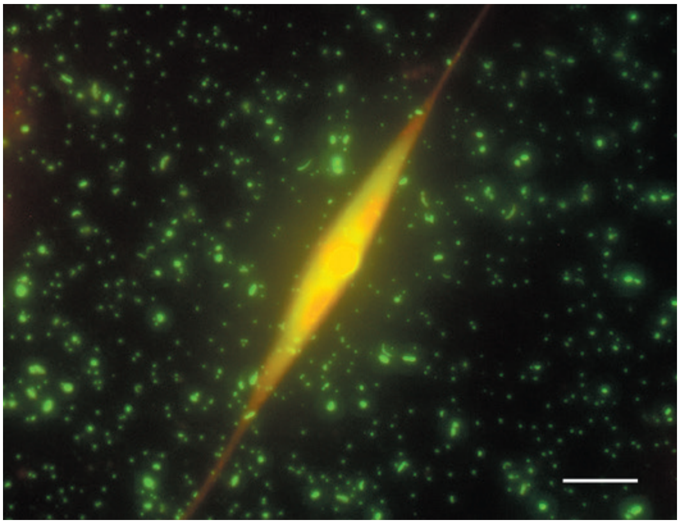
**The finding that marine environments with high levels of host microbes have fewer viruses per host than when host abundance is low challenges a theory on the relative roles of lysogenic and lytic viral-survival strategies.** SEE ARTICLE P.466

T. FREDE THINGSTAD & GUNNAR BRATBAK

Lysogeny is the process by which viral DNA is incorporated into the genome of the host organism, and it has long been thought that this ensures virus survival through periods of low host abundance. In this issue, Knowles *et al.*[1] (page 466) use viral and bacterial host-abundance and genomic data to suggest the opposite: that lysogeny is associated with high host abundances. The authors call this a 'piggyback-the-winner' strategy, as opposed to the conventional view, which they paraphrase as 'piggyback-the-loser'.

It is more than 25 years since the realization[2,3] that viruses are such abundant players in aquatic and other ecosystems that they are probably the most numerous biological entities on Earth[4] (Fig. 1). Over this time, virus ecology has grown into a key discipline of microbial ecology, opening our eyes to an amazingly diverse component of natural ecosystems[4,5]. Yet despite vast progress in the description of host–virus biomes, our understanding of the connections between host–virus systems and the ecosystems in which they are embedded remains vague.

The host–virus systems we observe today can be seen as the products of antagonistic arms races, in which both hosts and viruses have evolved strategies to ensure their survival and propagation not only during resource-rich periods of rapid growth, but also in leaner times. Like all specialized predators and parasites, virus lineages face the evolutionary dilemma that too much success is a potential disaster: an organism that drives its prey or hosts to extinction does not survive. This could be a particular challenge in variable environments in which 'too efficient' is a condition that changes continuously.



GUNNAR BRATBAK

**Figure 1 | Marine microbes and their viruses.** This fluorescence-microscopy image of marine microorganisms shows bacteria (oval or rod-shaped green bodies) and viruses (small dots) surrounding a diatom algal cell (yellow and orange). Aquatic environments typically have a ratio between free virus and potential hosts of around 10 to 1. Invisible with this technique are lysogenic viruses, whose viral DNA is incorporated into the host genome. Knowles *et al.*[1] present data that challenge conventional ideas of when and why lysogeny occurs in nature. (Scale bar, 20 micrometres.)

Phages — viruses that infect bacterial hosts — have developed two main life strategies, lysis and lysogeny. Lytic phages (also known as virulent phages) reproduce in their hosts to produce progeny that are released as the host cell bursts (lyses). Maintaining a population of lytic viruses requires that, on average, at least one of the viruses released in a lytic event finds and successfully infects a new host before the virus itself is inactivated. This requirement is obviously problematic if hosts become rare. Lysogenic phages (also known as temperate phages) incorporate their DNA into the host's genome, such that their DNA is copied together with that of the host cell as the cell grows and multiplies. Generations later, the phage DNA may reactivate to produce progeny viruses. Lysogeny has typically been argued to ensure not only that the phages survive without killing their hosts when hosts are few, but also that they are in the same place as the hosts when growth conditions improve. These arguments lead to an expectation of increased importance of lysogeny in oligotrophic (nutrient-poor) environments, as confirmed by some investigations[6].

The ecological consequences of the lytic and lysogenic strategies are different. By causing host-cell lysis, lytic viruses shunt energy and material out of the food chain, whereas lysogenic viruses have the potential to move genes between hosts in a process called transduction. The balance between lysis and lysogeny, and the environmental conditions that influence this balance, are therefore thought to have major consequences for how marine ecosystems work.

Knowles *et al.* present data that challenge the established view that low host abundance is a primary driver for a shift from lytic to lysogenic behaviour. When host abundance is high, the probability of a host–virus collision increases and one might expect an increasing role for lytic viruses, reflected in an increasing virus-to-microbe ratio. However, both in their own data from coral reefs and in a meta-analysis from a broad set of environments, Knowles *et al.* find the opposite: the ratio between virus and microbe abundance tends to decrease at high microbe abundances. They hypothesize that this is caused by an increasing tendency towards lysogeny at high host abundance.

However, other situations might produce a similar decreasing trend in the virus-to-microbe ratio. One example is a model[7] in which high-abundance-host communities are dominated by slow-growing, defensive host strains with low associated virus production, and in which viruses primarily attack more-competitive, fast-growing strains. The two models are not mutually exclusive, but Knowles and colleagues use metagenomic data (the combined genomes of an ecological community) to show that a set of DNA sequences associated with defence against viruses shows no correlation with microbial abundance, which argues against the defence hypothesis. By contrast, genes associated with the integration and excision of lysogenic viruses increase, supporting the authors' lysogeny hypothesis.

The metagenomic data sets now available[8] suggest that the gap between observed microbial diversity in the ocean and what we can explain theoretically is probably even larger than was recognized when this discrepancy was first pointed out more than 50 years ago[9]. Biodiversity theory thus remains one of the major challenges in marine microbial ecology. Viral lysis is thought to generate and maintain parts of host diversity[7], whereas lysogenic viruses could, theoretically, exist independently of diversity in the host community. Lytic and lysogenic viral strategies sit at the hub of a story that connects microbial diversity, activity, evolution and ecosystem function, yet the story is unfinished. Although lysogeny has so far played a modest part as a survival strategy, Knowles and colleagues' work suggests a more central role for this process in dynamic ecosystems. If confirmed by future work, this implies a reshuffling of key pieces of the puzzle that would have consequences for our understanding of the role of host–virus biology in structuring the microbial part of ocean ecosystems. ■

**T. Frede Thingstad** *and* **Gunnar Bratbak** *are in the Department of Biology, University of Bergen, 5020 Bergen, Norway.*
*e-mail: frede.thingstad@uib.no*

1. Knowles, B. *et al. Nature* **531,** 466–470 (2016).
2. Bergh, Ø., Børsheim, K. Y., Bratbak, G. & Heldal, M. *Nature* **340,** 467–468 (1989).
3. Proctor, L. M. & Fuhrman, J. A. *Nature* **343,** 60–62 (1990).
4. Suttle, C. A. *Nature Rev. Microbiol.* **5,** 801–812 (2007).
5. Rosariao, K. & Breitbart, M. *Curr. Opin. Virol.* **1,** 289–297 (2011).
6. Weinbauer, M. G. & Suttle, C. A. *Aquat. Microb. Ecol.* **18,** 217–225 (1999).
7. Thingstad, T. F., Våge, S., Storesund, J. E., Sandaa, R.-A. & Giske, J. *Proc. Natl Acad. Sci. USA* **111,** 7813–7818 (2014).
8. Sunagawa, S. *et al. Science* **348,** 1261447 (2015).
9. Hutchinson, G. E. *Am. Nat.* **95,** 137–145 (1961).

**This article was published online on 16 March 2016.**

PLANETARY SCIENCE

# Signs of a wandering Moon

**The presence of ice at two positions on opposite sides of the Moon suggests that the satellite's orientation was once shifted away from its present spin axis — a finding that has implications for the Moon's volcanic history.** SEE LETTER P.480

IAN GARRICK-BETHELL

The Moon's north and south poles are among the coldest regions in the Solar System — some areas are colder than Pluto. Water ice can remain stable here for billions of years, even when exposed to the vacuum of space. Scientists have detected small deposits of ice at the lunar poles; these have told us about how water moves across the lunar surface, but on page 480 of this issue, Siegler *et al.*[1] present measurements of lunar ice that tell us about events deep inside the Moon. These events cast light on the Moon's volcanic evolution and on its orientation in the sky. They also reveal that the face of the Moon seen today is different from the one that looked upon Earth billions of years ago.

Since at least 1961, scientists have speculated that volatile compounds such as water could slowly accumulate in the Moon's cold, shadowed polar craters[2]. The temperatures at these craters are so low because the Moon's spin axis is nearly perpendicular to the line connecting the Moon and the Sun (Fig. 1). Lunar seasons are therefore nearly non-existent, so the bottoms of some polar craters never see sunlight and hence stay cold enough to trap any water molecules that fall there. Unfortunately, the darkness in these craters also makes it hard to determine how much ice is in them. However, in 1998, scientists used an ingenious method to infer the presence of ice by observing how neutrons produced by space radiation interact with the hydrogen atoms in any near-surface water[3].

Using hydrogen data from more than a decade ago, Siegler *et al.* noticed that each lunar pole has a hydrogen deposit that is slightly displaced from the true north and south poles. Moreover, each of these off-axis deposits are symmetrically arranged, such that if the two poles are viewed on top of one another, a 180° rotation of the Moon would bring the deposits into perfect alignment. The authors therefore suggest that the Moon's orientation to its spin axis was once shifted away from the present one. This would have produced darkness, coldness and the accumulation of ice at these ancient 'palaeopoles'. The ice that formed in that era seems to have survived when the Moon eventually shifted to its present orientation, despite now being exposed to some sunlight (Fig. 1).

Siegler and co-workers strengthen their hypothesis by proposing a plausible mechanism for such a shift. Planets can change their orientation if their internal mass distribution changes. Pockets of dense material tend to be close to the equator to minimize the planet's spin energy. If a huge pile of lead weights suddenly appeared in New York, the city's latitude would eventually shift to a position slightly southward because of planetary reorientation. The opposite is also true — if New York suddenly became lower in density, it would shift northward. This process is known as true polar wander.

The authors therefore show that the observed shift in ice poles could be explained if certain regions on the Moon became lower in density. One of these regions, the Procellarum KREEP Terrane (PKT), is the Moon's most radioactive area, and was once the most volcanic — this volcanism was responsible for most of the dark plains of the Moon that can be seen from Earth. The radioactivity and volcanism imply that this region must have once been hot and thus less dense than its surroundings. The idea that true polar wander caused the

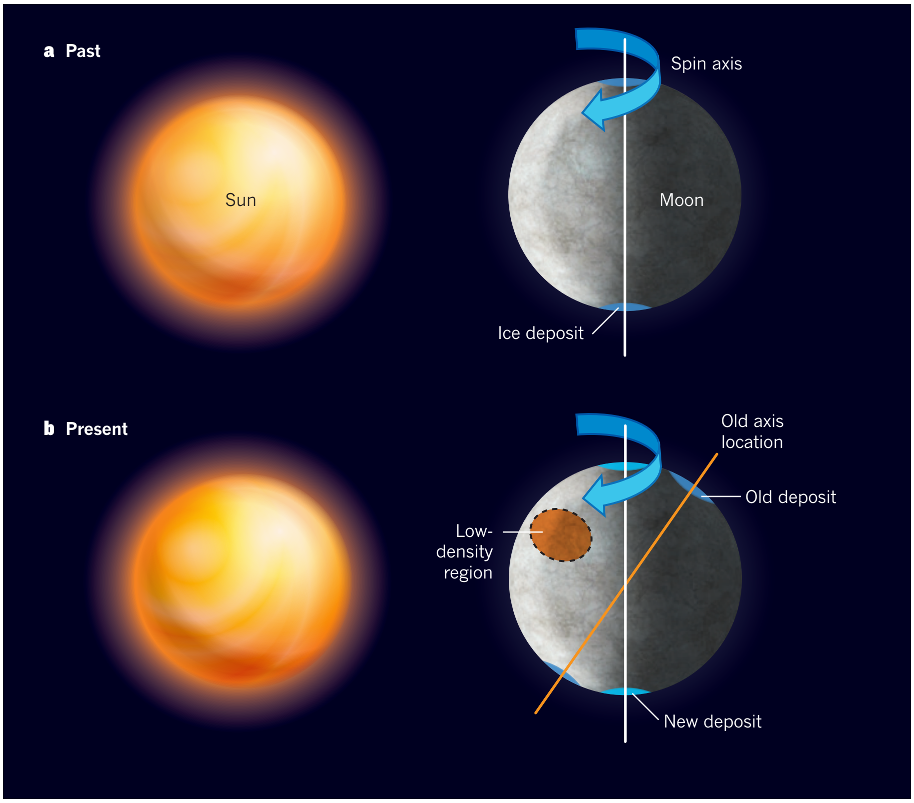


**Figure 1 | The Moon's shifting orientation.** Lunar seasons are almost non-existent because the Moon's spin axis is nearly perpendicular to the line that connects the Moon and the Sun. The bottoms of craters at the poles therefore never see sunlight and stay cold enough to trap water molecules. **a**, Siegler *et al.*[1] suggest that ice accumulated at the Moon's original poles in the past, forming two deposits. **b**, The authors propose that volcanic activity then generated an area of low density, which caused the Moon to reorient itself to minimize its spin energy. The original deposits have remained as evidence of the polar shift, and new ice deposits have formed at the current poles.

shifts therefore makes a lot of sense.

Siegler and colleagues' proposal has several implications for the history of lunar volcanism. If the PKT is indeed responsible for polar wander, it would help to constrain the magnitude of the heating events that occurred in that region. The scale of these events tells us how deep into the Moon the PKT reaches, and limits the timescales over which it became hot. This is important for understanding how the PKT became so radioactive in the first place — an enduring mystery in lunar science.

But it also presents a new puzzle. The Moon's volcanism mostly stopped 3 billion years ago, which means that the PKT has probably been getting colder and denser since then, not hotter. The direction of polar wander during this period would therefore have been in the opposite direction to the wander that produced the ice palaeopoles (see Fig. 4 of the paper[1]). Furthermore, one might expect ice palaeopoles to have formed everywhere along this polar-wander path, raising the question of why they are found only at the locations observed in the current study.

The authors' proposal also has implications for how long ago the ice deposits formed and how they have been preserved. The heating event that drove them to their present position began billions of years ago, implying that the deposits are similarly old. If so, how could they have remained there for so long, given that they are in sunlight? One explanation is that they must be buried below the equivalent of a permafrost level. Something similar happens on Earth — in Alaska, for example, temperatures below freezing occur a metre below the surface, even in summer.

But the Moon's soil has been churned over and over by countless asteroid impacts, which should have liberated some of this water. However, such impacts can also bury and protect ancient ice deposits. The details of the equilibrium between liberation and burial require further study, as does the exact nature of the ice and hydrogen at the poles.

Finally, as Siegler and co-workers acknowledge, their polar-wander hypothesis must be reconciled with other independent estimates of polar wander[4]. The amount of wander proposed by the authors is relatively modest, about 6°. Other work suggests that the Moon might have changed its orientation along a similar axis by 35° (ref. 5), or that much bigger changes occurred along multiple axes[6]. A key goal will be to reconcile these many stories of the changing orientation of the Moon, and to determine what density changes drove it to wander. ■

**Ian Garrick-Bethell** *is in the Department of Earth and Planetary Sciences, University of California, Santa Cruz, California 95064, USA.*
*e-mail: igarrick@ucsc.edu*

1. Siegler, M. A. *et al. Nature* **531,** 480–484 (2016).
2. Watson, K., Murray, B. & Brown, H. *J. Geophys. Res.* **66,** 1598–1600 (1961).
3. Feldman, W. C. *et al. Science* **281,** 1496–1500 (1998).
4. Runcorn, S. K. *Nature* **304,** 589–596 (1983).
5. Garrick-Bethell, I., Perera, V., Nimmo, F. & Zuber, M. T. *Nature* **512,** 181–184 (2014).
6. Takahashi, F., Tsunakawa, H., Shimizu, H., Shibuya, H. & Matsushima, M. *Nature Geosci.* **7,** 409–412 (2014).

BEHAVIOURAL ECONOMICS

# Corruption corrupts

**A cross-cultural experiment involving thousands of people worldwide shows that the prevalence of rule violations in a society, such as tax evasion and fraudulent politics, is detrimental to individuals' intrinsic honesty.** SEE LETTER P.496

SHAUL SHALVI

Does society affect intrinsic moral values? In this issue, Gächter and Schulz[1] (page 496) address this question with an experiment involving 2,568 participants in 23 countries. The authors show that a country's prevalence of rule violations, which for this study included tax evasion, corruption and political fraud, is positively associated with the tendency for residents of that country to lie for a small amount of extra cash. The finding rejects the idea that intrinsic honesty levels are similar in countries around the globe, and suggests that corruption corrupts.

Experimental research on human moral behaviour, for which intrinsic honesty is a proxy, is not overly concerned with how people 'should' behave. Instead, economists, psychologists and other researchers are descriptively mapping the situations in which people are likely to violate moral rules. The goal of such attempts is to craft useful interventions for encouraging moral conduct.

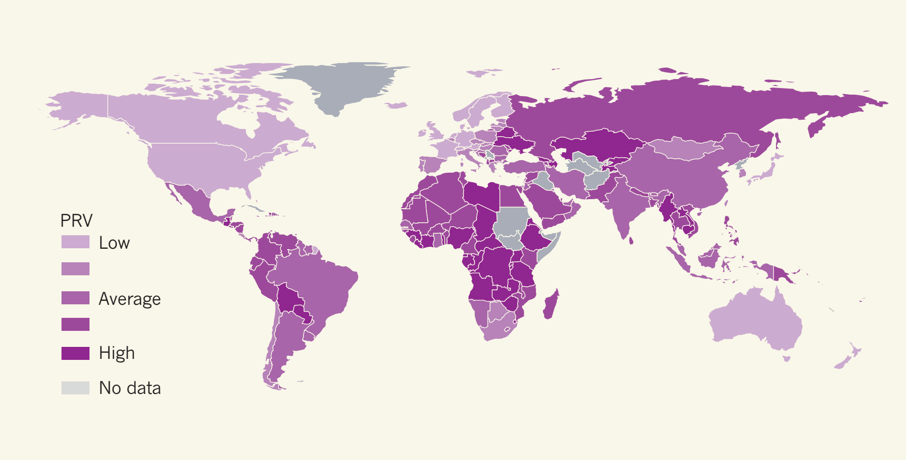


**Figure 1 | Rule violations across the globe.** Gächter and Schulz[1] developed a 'prevalence of rule violations' (PRV) index on the basis of a country's political democracy, illicit economic activity and levels of corruption. They assigned a PRV score to 159 countries, and investigated the effect of the relative prevalence of societal rule violations on individuals' honesty in 23 of those countries.

Indeed, people's deviant behaviour is influenced by their immediate environment. For example, people are more likely to drop litter, avoid returning their shopping trolleys and even trespass on private property when there are evident signs of disorder in their surroundings, such as graffiti[2]. But the extent to which corrupt societal norms trickle down to shape people's intrinsic standards of honesty remained unknown, until now. Tackling this fascinating issue, Gächter and Schulz used existing indices for the democratic quality of a country's political practices, its illicit economic activity and levels of corruption, to create a 'prevalence of rule violations' (PRV) index (Fig. 1).

The authors then used this index to classify 159 countries for which PRV-index data were available as of 2003, and investigated 23 representative countries. In each country, they sampled adult participants who were too young to have influenced the computed index. This is an essential ingredient in suggesting a causal path — that low exposure to rule violations increases people's intrinsic honesty, not vice versa.

Participants rolled a standard six-sided dice to determine their earnings in the experiment[3]. Operating in private, they rolled the dice, peeked at the outcome, then rolled and peeked a second time, and were asked to report the outcome of the first roll only. Higher reported numbers translated to higher earnings, with the exception of reporting a six, which meant getting nothing.

Because rolls were done in private, participants could easily misreport the outcome (lie) to increase their earnings. Although the task does not allow individual honesty or dishonesty to be pinpointed, the reports can be used to assess the degree and pattern of lying at the country level. In an honest country, given a large enough sample and a fair dice, the distribution of reported outcomes should be flat. The authors refer to this as the full-honesty benchmark. In a country in which people maximize profit at all costs, even by lying, only the most profitable value for the dice roll (five) should be reported — the full-dishonesty benchmark. Many people, however, like to feel moral even when lying, and thus prefer to shuffle facts rather than invent them. That is, people often report the higher observed outcome of the two rolls[4], not the value that appears on the first roll, as the rules dictate — the justified-dishonesty benchmark.

Gächter and Schulz found that participants were neither fully honest nor fully dishonest. Reported outcomes clustered around the justified-dishonesty benchmark, especially in countries with a high PRV score. This suggests that high exposure to rule violations turns people into truth stretchers, but not brazen liars. The authors also identify a positive correlation between a country's PRV score and participants' earnings in the task, suggesting that participants from more-corrupt countries lied more than those from less-corrupt ones. Given that participants were not involved in activities that could affect their country's score on the PRV index, the probable causal path is from society-level rule violations to individual-level dishonesty. Gächter and Schulz provide multiple tests that assess the robustness of the findings; for example, they show that use of the earliest available data related to PRV score, such as corruption levels in 1996, also predicts participants' dishonesty.

The underlying assumption of Gächter and Schulz's work is that country-level PRV score shapes country members' honesty, which is intrinsic and thus stable across situations. However, ample work suggests that the same person may be both honest and dishonest, according to situation[5–8]. For example, when people interact with a lying partner, they are likely to lie as well[9]. This elusive dynamic is missing when considering only snapshots of (dis)honesty.

***This suggests that high exposure to rule violations turns people into truth stretchers, but not brazen liars.***

Several intriguing questions remain open for future work. How long does it take for an individual's honesty to be shaped by their country's PRV score? According to a survey by Transparency International[10], corruption levels fell significantly in several countries, including Britain, Greece and Senegal, between 2012 and 2015. When should we expect to see more honesty in these countries? Furthermore, people are not confined to interacting with members of their own society. They travel abroad, do business internationally, attend student-exchange programmes and migrate. The impact of interacting with members of other countries on people's honesty remains an intriguing puzzle.

Most importantly, this study demonstrates that behavioural economic experimentation can provide insight into how to tackle burning global problems. A European Union anti-corruption report[11] estimated that corruption costs the EU €120 billion (US$132 billion) each year, just shy of its annual budget. The report concluded that "corruption seriously harms the economy and society as a whole". Gächter and Schulz's work makes it clear that the costs are not just financial. Corruption not only deprives people of economic prosperity and growth, but also jeopardizes their intrinsic honesty. ■

**Shaul Shalvi** *is in the Center for Research in Experimental Economics and Political Decision Making (CREED) and the Psychology Department, University of Amsterdam, 1018WB Amsterdam, the Netherlands.*
*e-mail: s.shalvi@uva.nl*

1. Gächter, S. & Schulz, J. F. *Nature* **531,** 496–499 (2016).
2. Keizer, K., Lindenberg, S. & Steg, L. *Science* **322,** 1681–1685 (2008).
3. Fischbacher, U. & Föllmi-Heusi, F. *J. Eur. Econ. Assoc.* **11,** 525–547 (2013).
4. Shalvi, S., Dana., J., Handgraaf, M. J. J. & De Dreu, C. K. W. *Organ. Behav. Hum. Decis. Processes* **115,** 181–190 (2011).
5. Shu, L. L., Mazar, N., Gino, F., Ariely, D. & Bazerman, M. H. *Proc. Natl Acad. Sci. USA* **109,** 15197–15200 (2012).
6. Cohn, A., Fehr, E. & Maréchal, M. A. *Nature* **516,** 86–89 (2014).
7. Schweitzer, M. E., Ordóñez, L. & Douma, B. *Acad. Mgmt J.* **47,** 422–432 (2004).
8. Maggian, V. & Villeval, M. C. *Exp. Econ.* http://dx.doi.org/10.1007/s10683-015-9459-7 (2015).
9. Weisel, O. & Shalvi, S. *Proc. Natl Acad. Sci. USA* **112,** 10651–10656 (2015).
10. Transparency International. www.transparency.org/cpi2015#results-table (2015).
11. European Commission. *EU Anti-Corruption Report*; available at go.nature.com/vsboih (2014).

**This article was published online on 9 March 2016.**

doi:10.1038/nature17396

# Direct synthesis of *Z*-alkenyl halides through catalytic cross-metathesis

Ming Joo Koh[1], Thach T. Nguyen[1], Hanmo Zhang[1], Richard R. Schrock[2] & Amir H. Hoveyda[1]

**Olefin metathesis has had a large impact on modern organic chemistry, but important shortcomings remain: for example, the lack of efficient processes that can be used to generate acyclic alkenyl halides. Halo-substituted ruthenium carbene complexes decompose rapidly or deliver low activity and/or minimal stereoselectivity, and our understanding of the corresponding high-oxidation-state systems is limited. Here we show that previously unknown halo-substituted molybdenum alkylidene species are exceptionally reactive and are able to participate in high-yielding olefin metathesis reactions that afford acyclic 1,2-disubstituted *Z*-alkenyl halides. Transformations are promoted by small amounts of a catalyst that is generated *in situ* and used with unpurified, commercially available and easy-to-handle liquid 1,2-dihaloethene reagents, and proceed to high conversion at ambient temperature within four hours. We obtain many alkenyl chlorides, bromides and fluorides in up to 91 per cent yield and complete *Z* selectivity. This method can be used to synthesize biologically active compounds readily and to perform site- and stereoselective fluorination of complex organic molecules.**

Olefins with a halogen substituent are a mainstay in chemistry. Alkenyl chlorides and bromides are found in biologically active natural products (for example, the recently isolated *Z*-alkenyl chloride containing neuromodulator janthielamide A[1] or bromine-containing fatty acids that are adipogenesis stimulators[2]) or can be used in some of the most central transformations in chemistry (for example, catalytic cross-coupling[3]). Alkenyl fluorides are valued because of the importance of organofluorine compounds in medicine[4], agrochemicals[5] and materials development[6]. A fluoro-substituted olefin can influence the properties of a molecule; an example is the *Z*-fluoroalkene derivative of $\gamma$-aminobutyric acid (GABA) transaminase inhibitor[7], which is more active than its *E* isomer[8] yet similarly potent and with a distinct mode of action compared to the parent non-fluorinated alkene (vigabatrin). Fluoro-olefins may be used as substrates in synthesis of fluorine-containing building blocks[9]. And yet the number of approaches for accessing alkenyl halides is limited; many entail multi-step sequences demanding prior synthesis of alkenylboron[10], alkenylsilane[11] or an organometallic species[12,13], followed by conversion of the C–B, C–Si or C–metal unit to a carbon–halogen bond (for a more extensive list, see Supplementary Information). Reactions might begin with the more costly and less widely available (compared to alkenes) alkyne substrates[13], at times proceed with moderate stereoselectivity[10], or are not sufficiently general[11,12]. Methods for preparation of 1,2-disubstituted *Z*-halo-alkenes with high stereoselectivity are even fewer in number[10–13]. One option is a Wittig reaction of an aldehyde with a halogen-substituted phosphonium salt[14,15], but stereoselectivities are variable and, at times, toxic hexamethylphosphoramide and/or severely low temperatures are needed for high *Z*:*E* ratios[16]. Approaches to synthesis of 1,2-disubstituted *Z*-alkenyl fluorides are scarce[15,17,18] and none has reasonable scope.

Certain 1,2-disubstituted *Z*-alkenyl halides can be prepared via stereodefined alkenyl–B(pin) (pin, pinacolato) compounds[19,20], accessible by catalytic cross-metathesis (CM) with vinyl–B(pin)[21,22]. If direct CM were able to deliver halogen-substituted olefins in a single catalytic reaction from a terminal olefin without the need for use and/or synthesis of (at times expensive) organoboron reagents, such a reaction would have several other advantages, as follows: (1) strong oxidants (for example, $Br_2$), toxic mercury salts[23] and/or the more difficult to prepare and use alkenylboronic acids[24] (compared to B(pin) derivatives) would not be needed; (2) severely basic conditions for (pin) B-to-halogen exchange and reactive halide sources (for example, iodine monochloride), which may be detrimental to certain functionality (for example, sulphides[25] or indoles[26]), would not be necessary; (3) product purification would be more practical—organic halide reagents are more easily removable (sufficiently volatile) and do not afford pinacol by-product that can be difficult to separate from the desired product; and (4) access to multifunctional molecules with an alkenyl–B(pin) as well as an alkenyl halide would be more feasible[27].

## The potential and challenge of alkenyl halide CM

A catalytic CM protocol that converts an alkene to an alkenyl halide directly would be complementary to the existing methods (it would offer a distinct synthesis strategy) and especially advantageous if a commercially available, easy-to-handle (that is, liquid at ambient conditions) and relatively inexpensive reagent could be used in a highly stereoselective process (Fig. 1a). For instance, a transition metal complex that catalyses CM of an abundant substrate such as methyl oleate and an easily accessible organo-chloride reagent would afford separable *Z*-alkenyl halide compounds (Fig. 1a); one (**1b**) could be converted to anti-inflammatory agent (*S*)-coriolic acid methyl ester[28] by an ensuing catalytic cross-coupling. Ring-opening/cross-metathesis (ROCM) of cyclooctene with an alkenyl bromide would deliver *Z*,*Z*-dibromoalkene **2**, an intermediate used to access anti-tumour and immunosuppressive agent tetrahydrosiphonodiol[29] (Fig. 1a). The feasibility of a CM that furnishes alkenyl fluorides would allow for late-stage fluorination[30] of complex molecules, such as potassium channel activator isopimaric acid[31] in a catalytic, chemo- and stereoselective fashion (**3**, Fig. 1a).

Development of efficient alkenyl halide-generating CM reactions is however not straightforward. Unlike Ru carbenes or Mo or W alkylidenes with alkyl, aryl, boryl or alkoxy substituents, those bearing a halogen atom are either unstable (Ru), their transformations inefficient (Ru)[32–34] or there is little known about them (Mo/W).

[1]Department of Chemistry, Merkert Chemistry Center, Boston College, Chestnut Hill, Massachusetts 02467, USA. [2]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

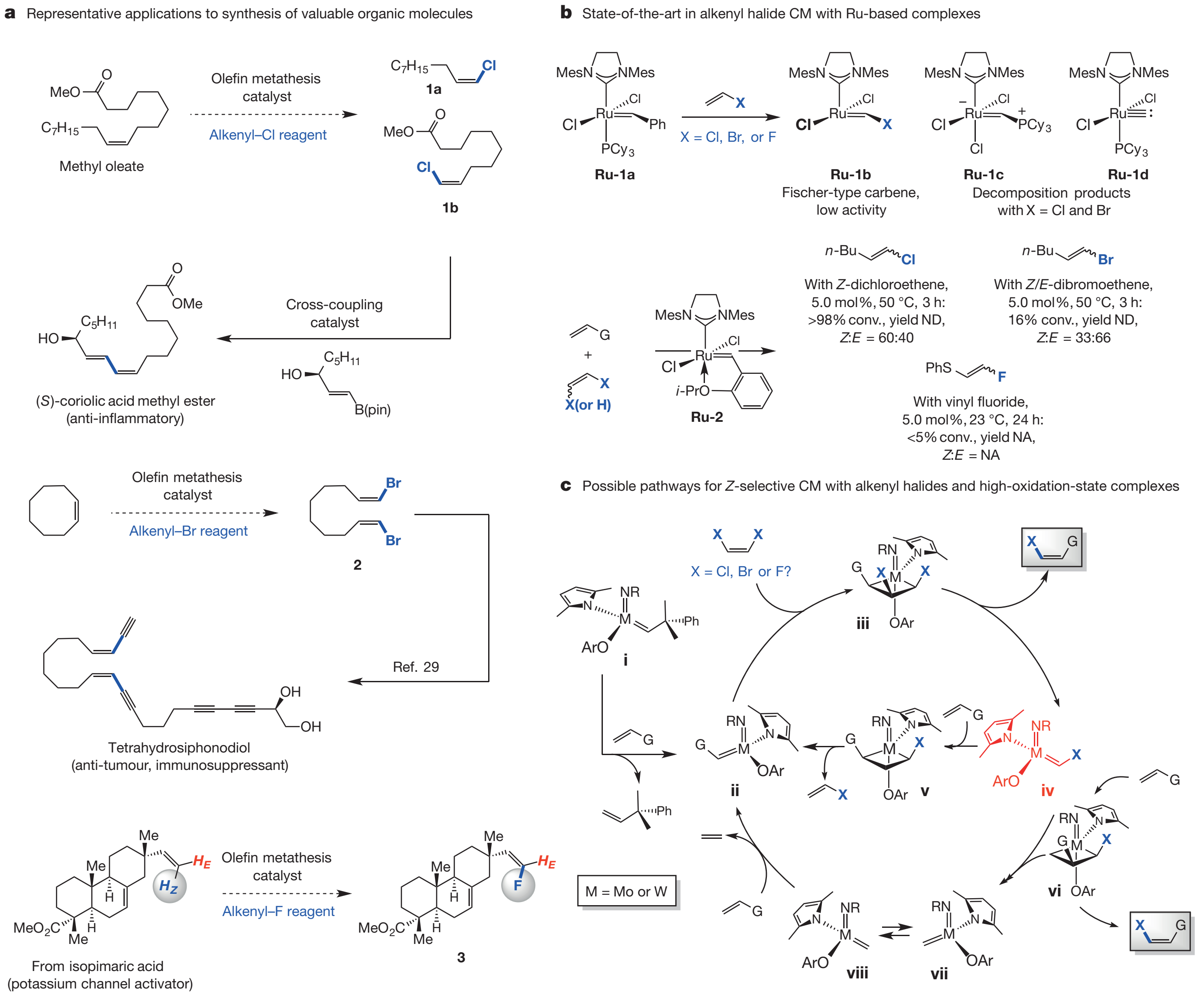

a Representative applications to synthesis of valuable organic molecules
MeO
O
$C_7H_{15}$
Methyl oleate
Olefin metathesis catalyst
Alkenyl–Cl reagent
$C_7H_{15}$
Cl
1a
MeO
O
Cl
1b
Cross-coupling catalyst
O
OMe
$C_5H_{11}$
HO
(S)-coriolic acid methyl ester (anti-inflammatory)
$C_5H_{11}$
HO
B(pin)
Olefin metathesis catalyst
Alkenyl–Br reagent
Br
Br
2
Ref. 29
OH
OH
Tetrahydrosiphonodiol (anti-tumour, immunosuppressant)
Me
Me
$H_E$
$H_Z$
H
H
$MeO_2C$
Me
From isopimaric acid (potassium channel activator)
Olefin metathesis catalyst
Alkenyl–F reagent
Me
Me
$H_E$
F
H
H
$MeO_2C$
Me
3
b State-of-the-art in alkenyl halide CM with Ru-based complexes
MesN
NMes
Cl
Cl
Ru
Ph
$PCy_3$
Ru-1a
X
X = Cl, Br, or F
MesN
NMes
Cl
Cl
Ru
X
Ru-1b
Fischer-type carbene, low activity
MesN
NMes
Cl
Ru
Cl
Cl
−
+
$PCy_3$
Ru-1c
MesN
NMes
Cl
Cl
Ru
$PCy_3$
Ru-1d
Decomposition products with X = Cl and Br
G
+
X
X(or H)
MesN
NMes
Cl
Cl
Ru
i-PrO
Ru-2
n-Bu
Cl
With Z-dichloroethene, 5.0 mol%, 50 °C, 3 h: >98% conv., yield ND, Z:E = 60:40
n-Bu
Br
With Z/E-dibromoethene, 5.0 mol%, 50 °C, 3 h: 16% conv., yield ND, Z:E = 33:66
PhS
F
With vinyl fluoride, 5.0 mol%, 23 °C, 24 h: <5% conv., yield NA, Z:E = NA
c Possible pathways for Z-selective CM with alkenyl halides and high-oxidation-state complexes
X
X
X = Cl, Br or F?
NR
N
M
Ph
ArO
i
RN
N
G
X
M
X
iii
OAr
X
G
G
RN
N
G
M
OAr
ii
RN
N
G
M
X
OAr
v
X
G
NR
N
M
X
ArO
iv
G
Ph
RN
N
G
M
X
vi
OAr
X
G
M = Mo or W
G
NR
N
M
ArO
viii
RN
N
M
OAr
vii

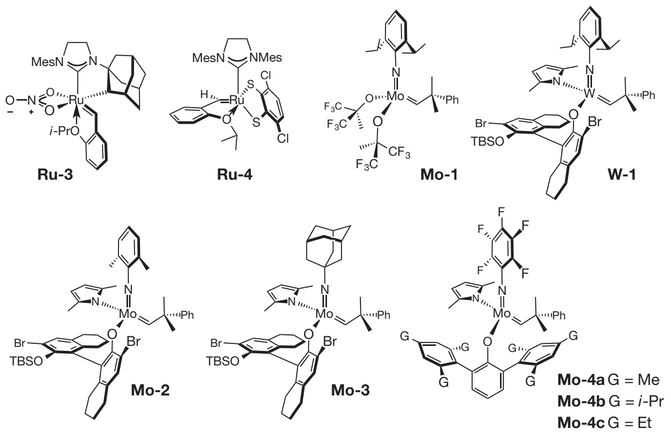
**Figure 1 | Designing catalytic CM reactions that afford *Z*-alkenyl halides. a**, Potential applications of *Z*-selective CM reactions that afford alkenyl halides include a concise synthesis of an anti-inflammatory agent (top) and a ring-opening/CM (ROCM) process that delivers a compound with two *Z*-alkenyl bromide bonds formerly employed in the preparation of an immunosuppressant (middle). The ability to perform late-stage stereoselective fluorination of complex molecules is another notable and high impact advantage (bottom). In each case, a readily accessible, easy-to-use and inexpensive halogen-containing reagent is involved, shown blue; alkenyl-Cl (top), alkenyl-Br (middle) and alkenyl-F (bottom). **b**, Ru complexes cannot promote efficient CM reactions of alkenyl halides because of the low reactivity and instability of the derived halo-substituted carbenes. **c**, The pathways that might allow a Mo or W species to promote CM transformations that generate alkenyl halides, despite the fleeting nature of the corresponding halo-substituted alkylidenes. Attempts to generate and examine Cl-substituted alkylidenes (**iv**, shown red, with X = Cl) were unsuccessful. Abbreviations: Conv., conversion; M, transition metal; X, halogen; Mes, 2,4,6-$(Me)_3C_6H_2$; G or R, various functional groups; Ar, aryl group; NA, not applicable; ND, not determined.

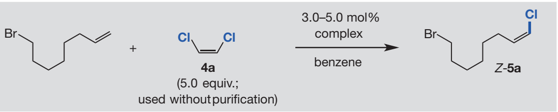
Fluoro-, chloro-, or bromo-substituted Fischer-type Ru complexes show negligible activity (**Ru-1b**, Fig. 1b)[34]. With phosphine-containing systems (for example, **Ru-1a**) inactive species such as phosphoniomethylidene **Ru-1c** and carbide **Ru-1d**[32] are produced. There is some improvement with phosphine-free complexes (for example, **Ru-2**, Fig. 1b)[33,34], but reactions are low yielding and minimally stereoselective despite elevated temperatures (for example, 50 °C) and long reaction times (for example, 24 h).

## Identification of an effective catalyst

The central issue, therefore, was whether high-oxidation-state (Mo/W) halo-substituted alkylidene complexes would be sufficiently robust yet appropriately reactive. Since alkoxy-substituted Mo alkylidenes are more active than the related Ru carbenes[35], we hoped that the same might apply to halogen-containing olefins, but we did not know of any data on the structure, stability or reactivity of a halo-substituted Mo or W alkylidene. Adding to the uncertainty is a computational study suggesting that fluoro-substituted Mo alkylidenes would be less stable than even the methylidenes[36]. Equally discouraging were the outcome of our attempts to prepare halo-substituted alkylidenes of Mo monoaryloxide pyrrolide (MAP) species (compare **iv**, Fig. 1c) by using *Z*-dichloroethene (**4a**). Subjection of a neophylidene MAP complex (compare **i**) with two equivalents of **4a** resulted in <2% transformation (4 h, 22 °C; 400 MHz $^1$H NMR analysis). The more reactive methylidene (generated from ethylene) was consumed completely, but a halo-substituted alkylidene was not found spectroscopically. Our remaining hope was that, although undetected, the putative complex might be sufficiently long-lived to fuel the catalytic cycles (compare **iv**, Fig. 1c). If so, reaction of a neophylidene with a terminal alkene could generate the less congested **ii**, which in turn might react with a

**Table 1 | Examination of complexes for CM**





| Entry number | Complex; loading (mol%) | Time (h); temperature (°C) | Conversion (%)*; yield (%)† | Z:E‡ |
|---|---|---|---|---|
| 1 | **Ru-2**; 5.0 | 4; 50 | 82; 59 | 58:42 |
| 2 | **Ru-3**; 5.0 | 4; 50 | 10; <5 | NA |
| 3 | **Ru-4**; 5.0 | 4; 50 | <10; <5 | NA |
| 4 | **Mo-1**; 5.0 | 4; 22 | 67; <5 | NA |
| 5 | **W-1**; 5.0 | 4; 22 | 45; <10 | ND |
| 6 | **Mo-2**; 5.0 | 4; 22 | 43; <5 | NA |
| 7 | **Mo-3**; 5.0 | 4; 22 | 60; 27 | >98:2 |
| 8 | **Mo-4a**; 5.0 | 4; 22 | 87; 60 | >98:2 |
| 9 | **Mo-4b**; 5.0 | 4; 22 | 62; 40 | 98:2 |
| 10 | **Mo-4b**; 5.0 | 12; 22 | 95; 84 | 93:7 |
| 11 | **Mo-4c**; 3.0 | 4; 22 | 90; 75 | >98:2 |

Here we show the performance of various complexes (shown above the table) used for CM of a terminal alkene with *Z*-1,2-dichloroethene (reaction highlighted in grey). Reactions were carried out under a nitrogen atmosphere; see Supplementary Information for details. NA, not applicable; ND, not determined.

*Conversion was based on the disappearance of the limiting reagent (8-bromo-1-octene) and determined by analysis of the $^1$H NMR spectra of the unpurified mixtures; the variance of values is estimated to be <±2%.

†Yield of isolated and purified product (*Z*/*E* mixture); the variance of values is estimated to be <±5%.

‡*Z*/*E* ratios were determined by $^1$H NMR analysis of unpurified mixtures; the variance of values is estimated to be <±2%. See Supplementary Information for details.

*Z*-dihaloalkene (compared to the more volatile vinyl halide), affording the desired product and halo-substituted alkylidene (**iv**) via all-*syn* metallacyclobutane **iii**. Complex **iv** and the olefin could then combine to afford **v**, which would in turn release alkylidene **ii** and vinyl halide. Alternatively, the halo-substituted alkylidene could react with another substrate molecule to furnish, by means of **vi**, the *Z*-alkenyl halide product and methylidenes **vii** and **viii**, which are precursors to **ii**.

We probed the ability of several complexes to effect *Z*-selective CM between 8-bromo-1-octene and commercially available and easy to handle *Z*-dichloroethene **4a** (boiling point 60 °C versus −13 °C for vinyl chloride). Reaction with dichloro complex **Ru-2** required 50 °C to reach 82% conversion after four hours (Table 1, entry 1), affording **5a** as a near equal mixture of stereoisomers; there was no transformation with *Z*-selective **Ru-3**[37] or **Ru-4**[38] (entries 2, 3). Use of bis-alkoxide **Mo-1** led to ~70% conversion (4 h, 22 °C) but mostly to the corresponding homocoupling product without any detectable alkenyl halide (entry 4). Experiments with complexes **W-1** and **Mo-2** were similarly disappointing (entries 5, 6) as again there was only alkene homocoupling (<2% **5a**). Adamantylimido **Mo-3** provided the first hopeful data: we isolated **5a** in 27% yield and >98% *Z* selectivity (entry 7). Efficiency improved with perfluoroimido complex **Mo-4a**: **Z-5a** was obtained in 60% yield with none of the alternative *E* isomer being observable ($^1$H NMR analysis, 4 h, 22 °C; entry 8). We then reasoned that a larger aryloxide ligand, although likely to be less active, might translate into longer catalyst lifetime and better efficiency; we therefore examined the CM with **Mo-4b**, but, while high stereochemical control could be retained (98:2 *Z*:*E*), conversion and yield were reduced (62% conversion, 40% yield; entry 9). After 12 h, **5a** was isolated in 84% yield (95% conversion; entry 10) but with some diminution in stereoisomeric purity (93:7 *Z*:*E*), probably caused by post-metathesis isomerization. To achieve a better balance between robustness and reaction rate without forfeiting stereocontrol, we examined 2,4,6-triethyl-substituted aryloxide complex **Mo-4c** (entry 11); **5a** could thus be secured in 75% yield and >98:2 *Z*:*E* selectivity after four hours at room temperature.

## Synthesis of *Z*-alkenyl chlorides and bromides

An array of *Z*-alkenyl chlorides can be prepared; yields were in the 50%–91% range with uniformly high stereoselectivity (95:5 to >98:2 *Z*:*E*; Fig. 2); the dichloroethene reagent (**4a**) was used without purification. Commonly occurring and versatile functional groups such as a silyl ether (**5b**, Fig. 2a), a sulphide (**5c**), an alkyne (**5d**), an epoxide (**5e**), an ester (**5f**) or a phthalimide (**5g**) were tolerated. An aryl or a heteroaryl moiety at the allylic position did not hinder the CM process (**5j**, **k**), but reactions with styrenes (regardless of its electronic attributes) were inefficient; this is probably due to steric hindrance within the requisite trisubstituted all-*syn* metallacyclobutane intermediate (compare **iii**, Fig. 1c) and the relatively facile homocoupling of aryl olefins. Hence, stilbenes, which do not re-enter the catalytic cycle easily (compared to the homocoupling product of an aliphatic alkene), were produced predominantly (see below for further discussion); however, in the reactions with α-branched aliphatic alkenes, which do not undergo homocoupling as rapidly for steric reasons, CM is efficient. *Z*-Selective synthesis of polycyclic compound **5n** demonstrates applicability to alkenes with a homoallylic quaternary carbon centre.

Allylboronate **5o** (Fig. 2b) was isolated in 66% yield and >98:2 *Z*:*E* selectivity; this product, similar to allyltin compound **5h** and allylsilane product **5i**, may be used as a reagent for C–C bond formation. Two representative cases are shown; in one, allyl chloride **6a** was obtained in >98% γ- and diastereoselectivity, and in the other, performed in the presence of 10 mol% aminophenol **7**[39], alkenyl chloride **6b** was generated with high α-selectivity without any loss in *Z*:*E* ratio (>98:2). As noted, access to several of the aforementioned CM products (such as sulphide **5c**, stannane **5h**, indole **5k** as well as allyl boron compound **5o**) by means of the two-step protocol involving vinyl–B(pin) CM/boron-to-halogen exchange would be problematic.

*Z*-Disubstituted alkenes are effective substrates. Treatment of commercially available *Z*-5-decene and *Z*-dichloroethene with 1.0 mol% **Mo-4c** for two hours followed by the addition of alkyne **8** (5.0 mol% $PdCl_2(PhCN)_2$, 10 mol% CuI, piperidine, 15 h), afforded **9** in 67% overall yield and 97:3 *Z*:*E* selectivity. These processes were performed without the need for isolation of volatile *Z*-alkenyl chloride **5p** (Fig. 2b), and the enyne product has been used in the synthesis of marine metabolite clathculin B[40]. Reactions can be easily carried out on a gram scale: CM of methyl oleate and **4a** in the presence of 3.0 mol% **Mo-4c** afforded *Z*-alkenyl chlorides **1a** and **1b** in 86% and 91% yield and with 97:3 *Z*:*E* selectivity, respectively (Fig. 2b). Subsequent catalytic cross-coupling with alkenylboronate **10**, obtained from site- and *E*-selective catalytic protoboryl addition of the commercially available propargyl alcohol[41], completed the two-step synthesis of (*S*)-coriolic acid methyl ester[28] from a renewable resource in 65% overall yield and 97:3 *Z*:*E* selectivity (compared to five steps previously; see Supplementary Information for bibliography).

*Z*-Selective synthesis of alkenyl bromides brings with it the added complication that stereoisomerically pure *Z*-dibromoethene (**4b**) is not readily available and difficult to prepare, but a 64:36 *Z*:*E* mixture can be purchased at relatively low cost (Fig. 3a). Although MAP complexes prefer to react with *Z*-1,2-disubstituted alkene isomers[42], our

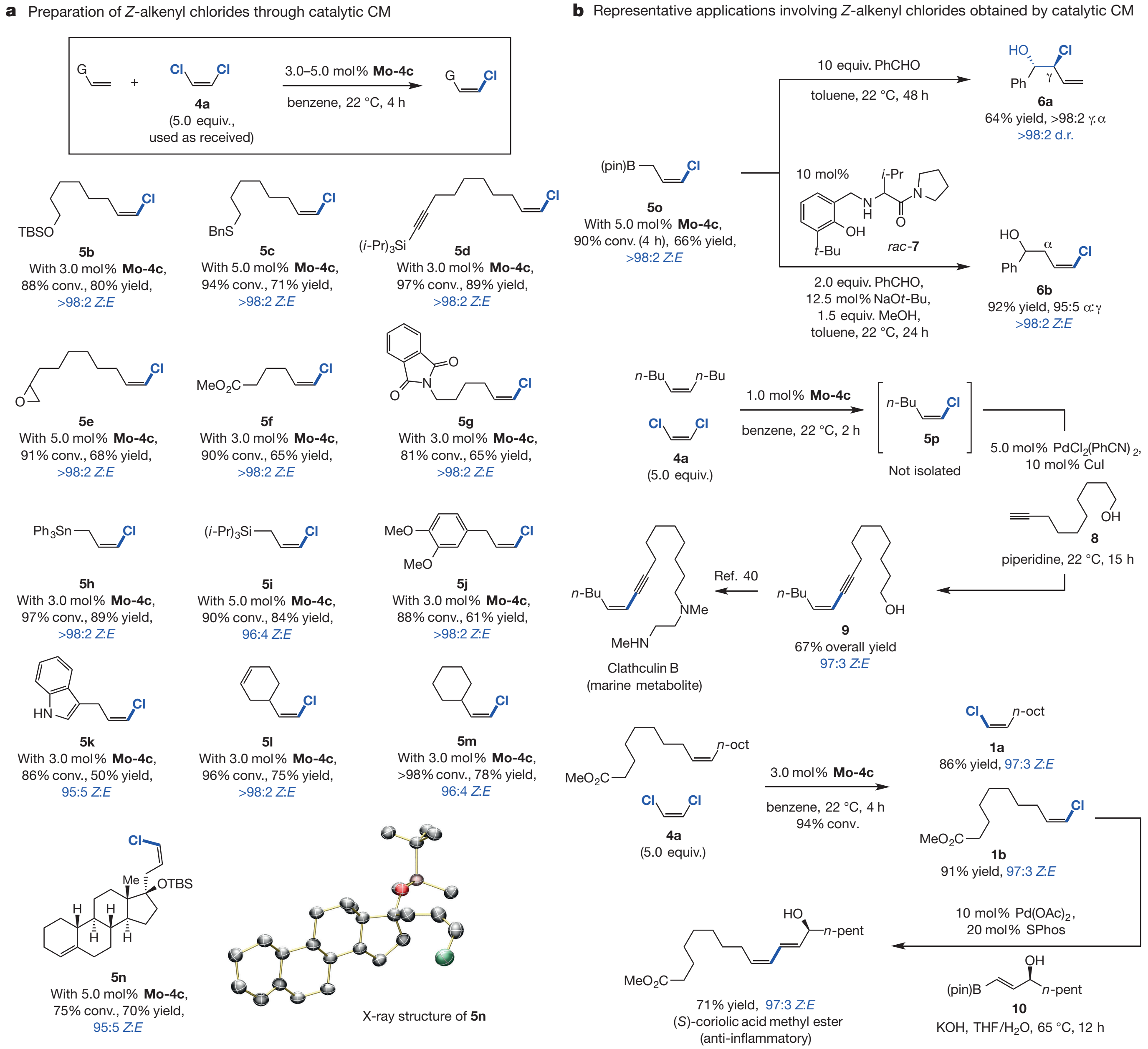

a Preparation of Z-alkenyl chlorides through catalytic CM
G
+
Cl Cl
4a
(5.0 equiv., used as received)
3.0–5.0 mol% Mo-4c
benzene, 22 °C, 4 h
G Cl
TBSO
5b
With 3.0 mol% Mo-4c, 88% conv., 80% yield, >98:2 Z:E
BnS
5c
With 5.0 mol% Mo-4c, 94% conv., 71% yield, >98:2 Z:E
(i-Pr)3Si
5d
With 3.0 mol% Mo-4c, 97% conv., 89% yield, >98:2 Z:E
5e
With 5.0 mol% Mo-4c, 91% conv., 68% yield, >98:2 Z:E
MeO2C
5f
With 3.0 mol% Mo-4c, 90% conv., 65% yield, >98:2 Z:E
5g
With 3.0 mol% Mo-4c, 81% conv., 65% yield, >98:2 Z:E
Ph3Sn
5h
With 3.0 mol% Mo-4c, 97% conv., 89% yield, >98:2 Z:E
(i-Pr)3Si
5i
With 5.0 mol% Mo-4c, 90% conv., 84% yield, 96:4 Z:E
MeO
MeO
5j
With 3.0 mol% Mo-4c, 88% conv., 61% yield, >98:2 Z:E
HN
5k
With 3.0 mol% Mo-4c, 86% conv., 50% yield, 95:5 Z:E
5l
With 3.0 mol% Mo-4c, 96% conv., 75% yield, >98:2 Z:E
5m
With 3.0 mol% Mo-4c, >98% conv., 78% yield, 96:4 Z:E
Me OTBS
H H H H
5n
With 5.0 mol% Mo-4c, 75% conv., 70% yield, 95:5 Z:E
X-ray structure of 5n
b Representative applications involving Z-alkenyl chlorides obtained by catalytic CM
(pin)B
5o
With 5.0 mol% Mo-4c, 90% conv. (4 h), 66% yield, >98:2 Z:E
10 equiv. PhCHO
toluene, 22 °C, 48 h
HO Cl Ph γ
6a
64% yield, >98:2 γ:α
>98:2 d.r.
10 mol%
i-Pr
N H
OH
t-Bu
rac-7
2.0 equiv. PhCHO, 12.5 mol% NaOt-Bu, 1.5 equiv. MeOH, toluene, 22 °C, 24 h
HO α Ph
6b
92% yield, 95:5 α:γ
>98:2 Z:E
n-Bu n-Bu
4a
(5.0 equiv.)
1.0 mol% Mo-4c
benzene, 22 °C, 2 h
n-Bu Cl
5p
Not isolated
5.0 mol% PdCl2(PhCN)2, 10 mol% CuI
OH
8
piperidine, 22 °C, 15 h
n-Bu
OH
9
67% overall yield
97:3 Z:E
Ref. 40
n-Bu
NMe
MeHN
Clathculin B
(marine metabolite)
MeO2C n-oct
4a
(5.0 equiv.)
3.0 mol% Mo-4c
benzene, 22 °C, 4 h
94% conv.
Cl n-oct
1a
86% yield, 97:3 Z:E
MeO2C
1b
91% yield, 97:3 Z:E
10 mol% Pd(OAc)2, 20 mol% SPhos
OH
(pin)B n-pent
10
KOH, THF/H2O, 65 °C, 12 h
HO n-pent
MeO2C
71% yield, 97:3 Z:E
(S)-coriolic acid methyl ester
(anti-inflammatory)

**Figure 2 | Synthesis of *Z*-alkenyl chlorides and applications.** **a**, Many *Z*-alkenyl chlorides (**5b**–**5n**) can be prepared with **Mo-4c** and unpurified *Z*-dichloroethene (**4a**; see boxed reaction at top); each product **5b**–**5n** has synthesis conditions shown under. Useful functional units are tolerated, among them a sulphide, an allyl stannane, an indole and an allylboron. The X-ray structure of **5n** confirms the predominant formation of the *Z* isomer. **b**, Chloro-substituted allylboron compounds for use in catalytic C–C bond forming transformations. Top, synthesis of **6a**, **b**. Application to synthesis of clathculin B (middle) and (*S*)-coriolic acid methyl ester (bottom, on a one gram scale) further underscores utility. Abbreviations: G, functional groups; TBS, *t*-butyldimethylsilyl; Bn, benzyl; pin, pinacolato; Ac, acetyl; SPhos, 2-dicyclohexylphosphino-2′,6′-dimethoxybiphenyl. Reactions were performed under $N_2$. Conversions and *Z*:*E* ratios were measured by analysis of $^1$H NMR spectra of unpurified mixtures; the variance of values is estimated to be <±2%. Yields correspond to isolated and purified products and represent an average of at least three runs (±5%). See Supplementary Information for experimental details and spectroscopic analyses.

concern was possible interference by *E*-**4b**, leading to diminution in stereoselectivity. It was also unclear whether the more sizeable dibromoethene would cause significant lowering of efficiency. In the event, a range of *Z*-alkenyl bromides were obtained in 57%–83% yield and 87:13–91:9 *Z*:*E* selectivity (**11a**–**f**, Fig. 3a). With the more volatile vinyl bromide (versus **4b**), yields were significantly lower (<25%) because the increased amount of ethylene boosts the concentration of the comparatively unstable methylidene complexes (compare **vii**, **viii**, Fig. 1c). The lower *Z* selectivity in the case of bromoalkene products (versus alkenyl chlorides) may be attributed to a minor pathway involving metallacyclobutanes derived from the *E* isomer of the dibromo-alkene reagent (see Supplementary Information for details).

The present strategies are applicable to ROCM; two instances are depicted in Fig. 3b. Dibromoalkene **2** was obtained in 88% yield and 89:11 *Z*,*Z*:*Z*,*E* selectivity (10 mol% **Mo-4c**, 1 h); as mentioned (compare Fig. 1a), diene **2** has been used in the preparation of tetrahydrosiphonodiol[29]. The need for larger amounts of the more active pentafluoroimido **Mo-4c** is so that maximum amounts of the ring-opening polymerization (ROMP) by-product can be converted to monomeric **2**. *Z*,*Z*-Dichloroalkene **12** was isolated in 75% yield as a single stereoisomer; adamantylimido complex **Mo-3** proved optimal, as this less active catalyst (compared to **Mo-4c**) is sufficient for the faster ROCM involving the less hindered *Z*-dichloroethene to compete with ROMP for attaining maximal *Z* selectivity. When the milder **Mo-3** was

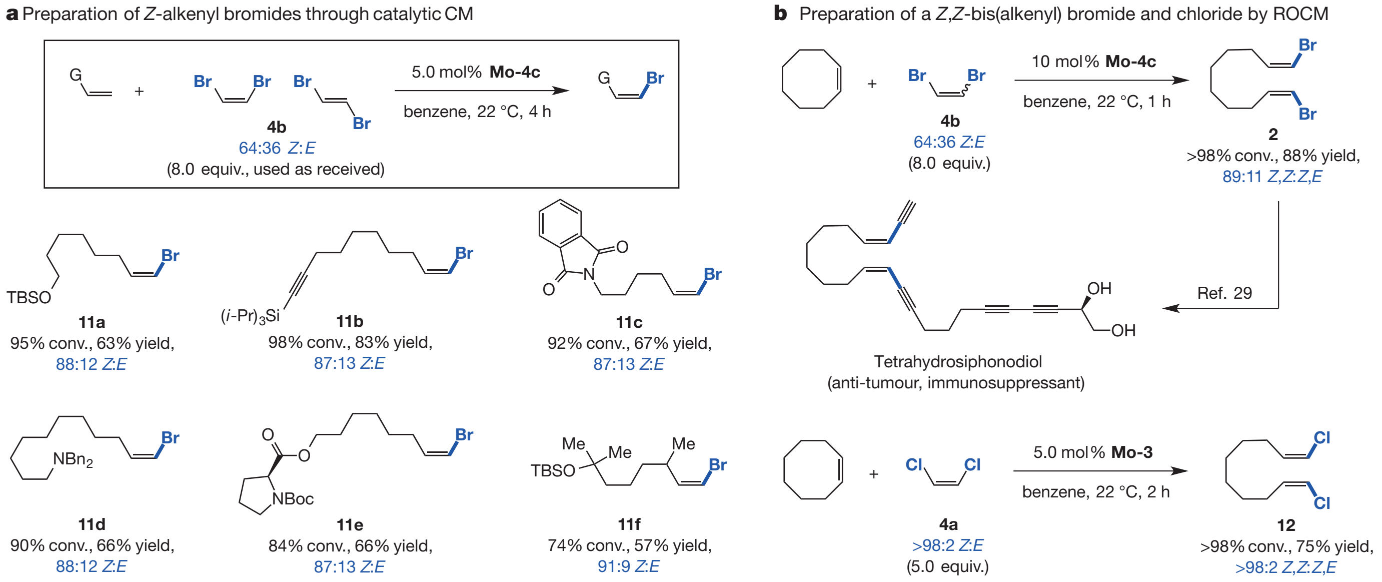


**Figure 3 | *Z*-Alkenyl bromides through catalytic CM and ROCM.** **a**, Stereoisomeric mixture of 1,2-dibromoethene (**4b**) can be used in preparation of *Z*-alkenyl bromides **11a**–**11f** via the reaction shown boxed at the top. **b**, The protocol is applicable to ROCM processes with readily accessible cyclic alkenes; *Z*,*Z*-bis(alkenyl)bromide has been employed in the preparation of anti-tumour agent tetrahydrosiphonodiol (top). The corresponding dichloride was synthesized in 75% yield and with complete *Z* selectivity (bottom). Abbreviations: G, various functional groups; Ar, aryl group; TBS, *t*-butyldimethylsilyl; Bn, benzyl; Boc, *t*-butyloxycarbonyl. Reactions were performed under $N_2$. Conversions and *Z*:*E* ratios were measured as in Fig. 2, with the same variance; yields were also measured as in Fig. 2. See Supplementary Information for experimental details and spectroscopic analyses.

used in the more demanding transformation leading to bromo-alkene **2**, there was >98:2 *Z*,*Z*:*Z*,*E* selectivity but with less conversion to the desired product (~35%, ~20% ROMP). Control experiments indicated that post-metathesis isomerization is minimal.

## Synthesis of *Z*-alkenyl fluorides

Development of *Z*-selective CM reactions that afford organofluorine products posed a new complication (Fig. 4). Vinyl fluoride has a very low boiling point (−72 °C versus −13 °C for vinyl chloride); *Z*-difluoroethene is expensive, similarly difficult to handle as well as explosive (Fig. 4a). We thus envisioned using *Z*-bromo-fluoroethene (**4c**), a commercially available, economically viable and substantially less volatile organohalide (boiling point, +36 °C), an option that raises a selectivity problem: the bromo-fluoroethene compound must interact with a Mo alkylidene according to the regiochemical mode of addition **I** in Fig. 4a. If the transformation were to proceed through **II**, a *Z*-alkenyl bromide would be formed. We reasoned that reaction via **I** might be preferred for two reasons. First, the $^1$H NMR spectrum ($CDCl_3$) of **4c** contains a significantly more upfield signal for the proton at the base of the C–Br bond, indicating that electron density is greater at this carbon (stronger π donation and σ withdrawing inductive effect by fluorine), favouring its association with the Lewis acidic Mo centre (compare **I** versus **II**). Additionally, the metallacyclobutane generated via **II** would suffer from steric repulsion between the more sizeable halogen and the alkylidene substituent (G). The catalytic CM affording *Z*-alkenyl fluoride **13a** indeed generated bromide **11b** as the minor product (72:28 fluoro:bromo; Fig. 4b). Consistent with the suggested model (**I** versus **II**), with an α-branched terminal alkene, the product mixture was less contaminated by the corresponding bromoalkene: pure **13b**, formed from a CM reaction that proceeded with 96:4 fluoro:bromo selectivity, was isolated in 70% yield and >98:2 *Z*:*E* ratio after purification.

Contrary to transformations of styrenes with dichloro- or dibromoethene (**4a**, **b**), CM with **4c** and aryl olefins proceeded readily and stereoselectively: β-(*Z*)-fluorostyrenes **13c**–**f** were obtained in 93:7–96:4 fluoro:bromo selectivity, 64%–72% yield of the pure *Z*-alkenyl fluoride and 93:7–97:3 *Z*:*E* selectivity. These variations in efficiency might be associated with the lower steric repulsion (eclipsing interaction of fluorine with G in the all-*syn* metallacyclobutane) versus the larger chlorine and bromine atoms, such that CM with **4c** competes better with homocoupling of styrene. To the best of our knowledge, there are no reports regarding the synthesis of aryl-substituted *Z*-alkenyl fluorides by catalytic cross-coupling of **4c**, and such transformations (for example, **13e**, **f**) would probably suffer from chemoselectivity complications. The present processes would offer an attractive pathway for accessing a variety of organofluorine compounds[43]. *Z*-Alkenyl fluoride **13g** has been converted to the aforementioned GABA transaminase inhibitor **14**[7]; product **13g** was obtained in 55% overall yield and >98:2 fluoro:bromo and *Z*:*E* selectivity by CM with the silyl-amide substrate followed by deprotection. There was <5% conversion with the parent amide probably due to internal association of the Lewis basic amide with the Mo centre in the intermediate alkylidene complex[44].

## *Z*-selective complex molecule fluorination

A corollary to the present approach is the possibility of implementing net stereoselective olefinic C–H/C–F bond exchange within a complex molecule; this would allow rapid access and screening of well-defined fluorine-tagged derivatives for possible desirable properties. In this context (Fig. 4c), formation of *Z*-alkenyl fluoride **15** (>98:2 fluoro:bromo, 63% yield, >98:2 *Z*:*E*) demonstrates relevance to processes involving a relatively hindered allylic ether[45]. Tricyclic product **3** (94:6 fluoro:bromo, 70% yield in the pure form, 96:4 *Z*:*E*) is derived from the challenging CM with the isopimaric acid[31] methyl ester (compare Fig. 1a); here, the alkene is next to a sterically demanding all-carbon quaternary centre.

The findings summarized in Fig. 4d illustrate that the method is tolerant of a range of functional units commonly found in biologically active molecules. *Z*-alkenyl fluoride **16** (from anti-depressant perphenazine[46]) was obtained efficiently and stereoselectively (91:9 fluoro:bromo, 78% yield, >98:2 *Z*:*E*), underscoring tolerance towards aryl or alkyl amines and aryl sulphides. Synthesis of *Z*-fluoro-alkene **17** (from β-lactamase inhibitor sulbactam[47]) by the two-step sequence of *Z*-selective CM with vinyl–B(pin)[21] followed by conversion of the C–B unit to a C–F bond, according to the only available reported procedure[48], led to outright substrate decomposition. The first step afforded the *Z*-alkenyl–B(pin) compound as expected (22 °C, 24 h, 70%

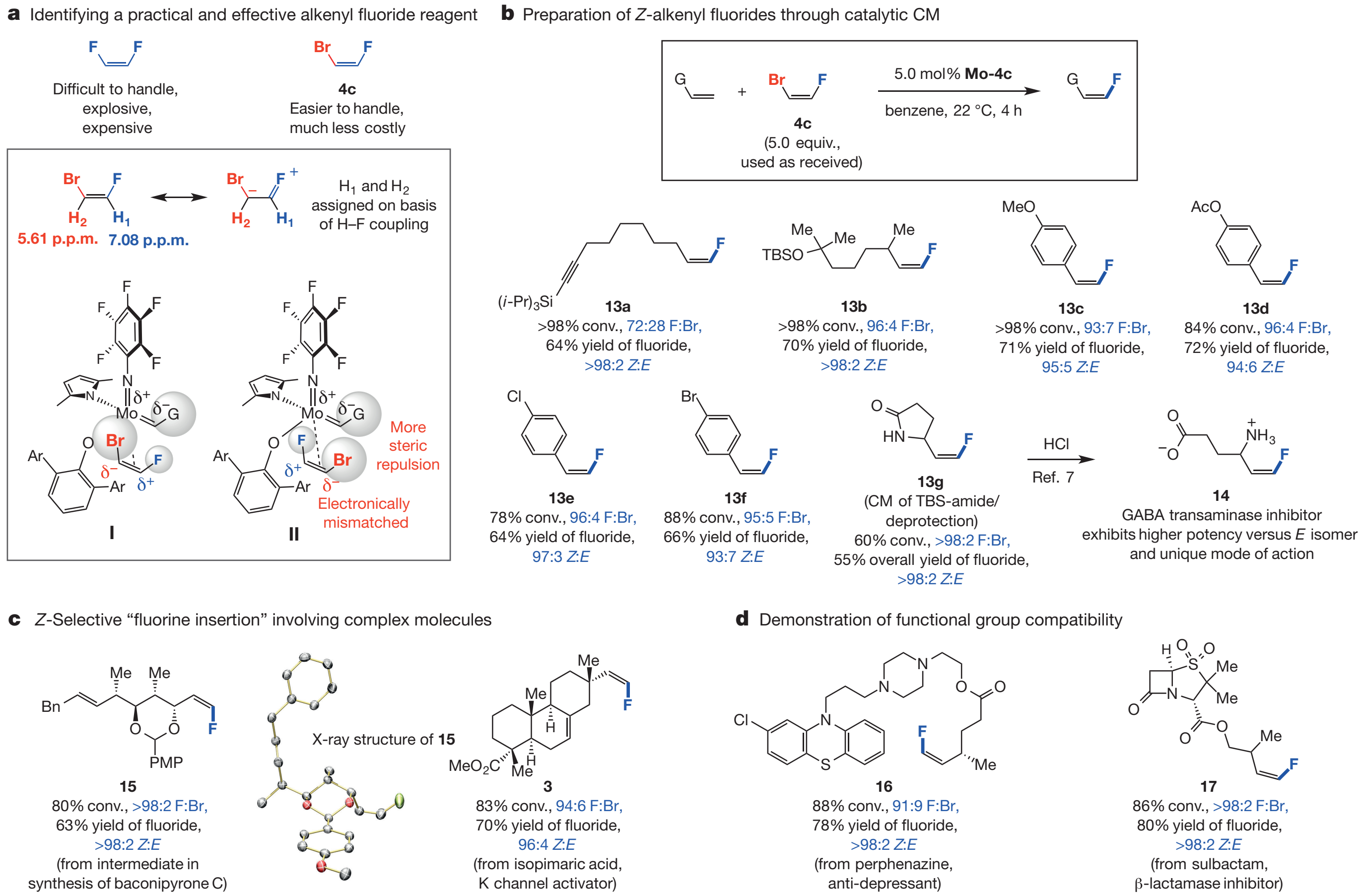

a Identifying a practical and effective alkenyl fluoride reagent
Difficult to handle, explosive, expensive
4c
Easier to handle, much less costly
$H_1$ and $H_2$ assigned on basis of H–F coupling
5.61 p.p.m.
7.08 p.p.m.
More steric repulsion
Electronically mismatched
I
II
b Preparation of Z-alkenyl fluorides through catalytic CM
5.0 mol% Mo-4c
benzene, 22 °C, 4 h
4c
(5.0 equiv., used as received)
13a
>98% conv., 72:28 F:Br, 64% yield of fluoride, >98:2 Z:E
13b
>98% conv., 96:4 F:Br, 70% yield of fluoride, >98:2 Z:E
13c
>98% conv., 93:7 F:Br, 71% yield of fluoride, 95:5 Z:E
13d
84% conv., 96:4 F:Br, 72% yield of fluoride, 94:6 Z:E
13e
78% conv., 96:4 F:Br, 64% yield of fluoride, 97:3 Z:E
13f
88% conv., 95:5 F:Br, 66% yield of fluoride, 93:7 Z:E
13g
(CM of TBS-amide/ deprotection)
60% conv., >98:2 F:Br, 55% overall yield of fluoride, >98:2 Z:E
HCl
Ref. 7
14
GABA transaminase inhibitor exhibits higher potency versus E isomer and unique mode of action
c Z-Selective "fluorine insertion" involving complex molecules
X-ray structure of 15
15
80% conv., >98:2 F:Br, 63% yield of fluoride, >98:2 Z:E
(from intermediate in synthesis of baconipyrone C)
3
83% conv., 94:6 F:Br, 70% yield of fluoride, 96:4 Z:E
(from isopimaric acid, K channel activator)
d Demonstration of functional group compatibility
16
88% conv., 91:9 F:Br, 78% yield of fluoride, >98:2 Z:E
(from perphenazine, anti-depressant)
17
86% conv., >98:2 F:Br, 80% yield of fluoride, >98:2 Z:E
(from sulbactam, β-lactamase inhibitor)

**Figure 4 | Z-alkenyl fluorides and late-stage fluorination. a**, *Z*-Bromo-fluoroethene (**4c**, top) can be used for synthesis of *Z*-alkenyl fluorides. Based on electronic and steric factors, reactions probably proceed via **I** (versus **II**). **b**, An array of products (**13a**–**14**) can be accessed using the general reaction shown boxed, including those with an aryl substituent. **c**, Stereoselective late-stage fluorination of complex molecules can be performed, giving, for example **15**, **3**. **d**, A variety of widely occurring heteroatom-containing functional units are tolerated, giving, for example, **16**, **17**. Abbreviations: G, various functional groups; Ar, aryl group; PMP, *p*-methoxyphenyl; Ac, acetyl; Bn, benzyl; TBS, *t*-butyldimethylsilyl. Reactions were performed under $N_2$. Conversions and *Z*:*E* ratios were measured as in Fig. 2, with the same variance; yields were also measured as in Fig. 2. For **13a**, 3.0 mol% **Mo-4c** was used, and for **3**, **13g** and **15**, 10 mol% **Mo-4c** was used (40 °C, 12 h for **13g**). See Supplementary Information for details.

conversion, >98:2 *Z*:*E*); but attempts to generate **17** by treatment with NaOH and AgOTf and then Selectfluor yielded an unidentifiable mixture of compounds, probably due to sensitivity of the substrate's bicyclic core[49]. In contrast, *Z*-alkenyl fluoride **17** was obtained through direct CM in 80% yield (>98:2 fluoro:bromo) as a single stereoisomer (>98% *Z*).

## Conclusions

We have introduced halo-substituted Mo alkylidenes as highly reactive and difficult-to-detect but viable intermediates in olefin metathesis. The matter of efficiency is especially noteworthy because, regardless of stereochemical control, there did not exist previously a catalytic CM protocol that generated halo-alkenes in useful yields. The ability of MAP catalysts to provide a solution to this central problem lies in their distinct electronic attributes, striking a balance between high reactivity and sufficient longevity. The catalytically active halo-substituted alkylidenes derived from **Mo-4c** can thus deliver the necessary activity (for example, versus **Ru-2**–**4** or **W-1**) but not at the expense of catalyst lifetime (for example, versus bis(alkoxide) **Mo-1**). The Mo centre in a MAP system is probably electron-deficient enough to prevent metal-carbide formation[34], and yet, unlike Ru carbenes, π-electron donation by a halide alkylidene substituent[50] does not hamper reactivity. Another noteworthy aspect is the design of reactions where the use of a dissymmetric *Z*-bromo-fluoroethene leads to the predominant or exclusive formation of fluoro-substituted alkenes (versus the bromo derivatives); this way, an easy-to-handle and readily accessible reagent can be used instead of the costly and impractical fluoro-olefin alternatives (for example, vinyl fluoride or *Z*-1,2-difluoroethene).

The advances outlined here serve as the foundation for future progress involving this intriguing set of halogen-containing Mo alkylidenes. The transformations should facilitate considerably the preparation of an assortment of desirable molecules for research in chemistry, biology and medicine, particularly since easy-to-handle (no glove box needed) paraffin-wrapped MAP complexes are becoming commercially accessible.

**Received 10 December 2015; accepted 2 February 2016.**

1. Nunnery, J. K. *et al.* Biosynthetically intriguing chlorinated lipophilic metabolites from geographically distant tropical marine cyanobacteria. *J. Org. Chem.* **77,** 4198–4208 (2012).
2. Akiyama, T. *et al.* Stimulators of adipogenesis from the marine sponge *Xestospongia testudinaria*. *Tetrahedron* **69,** 6560–6564 (2013).
3. Johansson Seechurn, C. C. C., Kitching, M. O., Colacot, T. J. & Sniekus, V. Palladium-catalyzed cross-coupling: a historical contextual perspective to the 2010 Nobel Prize. *Angew. Chem. Int. Ed.* **51,** 5062–5085 (2012).
4. Gillis, E. P., Eastman, K. J., Hill, M. D., Donnelly, D. J. & Meanwell, N. A. Applications of fluorine in medicinal chemistry. *J. Med. Chem.* **58,** 8315–8359 (2015).
5. Fujiwara, T. & O'Hagan, D. Successful fluorine-containing herbicide agrochemicals. *J. Fluor. Chem.* **167,** 16–29 (2014).
6. Berger, R., Resnati, G., Metrangolo, P., Weber, E. & Hulliger, J. Organic fluorine compounds: a great opportunity for enhanced materials properties. *Chem. Soc. Rev.* **40,** 3496–3508 (2011).

7. Kolb, M., Barth, J., Heydt, J.-G. & Jung, M. J. Synthesis and evaluation of mono-, di-, and trifluoroethenyl-GABA derivatives as GABA-T inhibitors. *J. Med. Chem.* **30,** 267–272 (1987).
8. Silverman, R. B., Bichler, K. A. & Leon, A. J. Unusual mechanistic difference in the activation of $\gamma$-aminobutyric acid aminotransferase by (*E*)- and (*Z*)-4-amino-6-fluoro-5-hexenoic acid. *J. Am. Chem. Soc.* **118,** 1253–1261 (1996).
9. Rosen, T. C., Yoshida, S., Kirk, K. L. & Haufe, G. Fluorinated phenylcyclopropylamines as inhibitors of monoamine oxidases. *ChemBioChem* **5,** 1033–1043 (2004).
10. Morrill, C. & Grubbs, R. H. Synthesis of functionalized vinylboronates via ruthenium-catalyzed cross-metathesis and subsequent conversion to vinyl halides. *J. Org. Chem.* **68,** 6031–6034 (2003).
11. Pawluć, P., Hreczycho, G., Szudkowska, J., Kubicki, M. & Marciniec, B. New one-pot synthesis of (*E*)-$\beta$-aryl vinyl halides from styrenes. *Org. Lett.* **11,** 3390–3393 (2009).
12. Bull, J. A., Mousseau, J. J. & Charette, A. B. Convenient one-pot synthesis of (*E*)-$\beta$-aryl vinyl halides from benzyl bromides and dihalomethanes. *Org. Lett.* **10,** 5485–5488 (2008).
13. Gao, F. & Hoveyda, A. H. $\alpha$-Selective Ni-catalyzed hydroalumination of aryl- and alkyl-substituted terminal alkynes: practical syntheses of internal vinyl aluminums, halides, or boronates. *J. Am. Chem. Soc.* **132,** 10961–10963 (2010).
14. Stork, G. & Zhao, K. A stereoselective synthesis of (*Z*)-1-iodo-1-alkenes. *Tetrahedr. Lett.* **30,** 2173–2174 (1989).
15. Zhang, X.-P. & Schlosser, M. Highly *cis*-selective Wittig reactions employing $\alpha$-heterosubstituted ylids. *Tetrahedr. Lett.* **34,** 1925–1928 (1993).
16. Crane, E. A., Zabawa, T. P., Farmer, R. L. & Scheidt, K. A. Enantioselective synthesis of (–)-exiguolide by iterative stereoselective dioxinone-directed Prins cyclizations. *Angew. Chem. Int. Ed.* **50,** 9112–9115 (2011).
17. Landelle, G., Turcotte-Savard, M.-C., Angers, L. & Paquin, J. F. Stereoselective synthesis of both stereoisomers of $\beta$-fluorostyrene derivatives from a common intermediate. *Org. Lett.* **13,** 1568–1571 (2011).
18. Nakagawa, M., Saito, A., Soga, A., Yamamoto, N. & Taguchi, T. Chromium mediated stereoselective synthesis of (*Z*)-1-fluoro-2-alkenyl alkyl and trialkylsilyl ethers from dibromofluoromethylcarbinyl ethers. *Tetrahedr. Lett.* **46,** 5257–5261 (2005).
19. Ohmura, T., Yamamoto, Y. & Miyaura, N. Rhodium- or iridium-catalyzed *trans*-hydroboration of terminal alkynes, giving (*Z*)-1-alkenylboron compounds. *J. Am. Chem. Soc.* **122,** 4990–4991 (2000).
20. Molander, G. A. & Ellis, N. M. Highly stereoselective synthesis of *cis*-alkenyl pinacolatoboronates and potassium *cis*-alkenyltrifluoroborates via a hydroboration/protodeboronation approach. *J. Org. Chem.* **73,** 6841–6844 (2008).
21. Kiesewetter, E. T. *et al.* Synthesis of *Z*-(pinacolato)allylboron and *Z*-(pinacolato) alkenylboron compounds through stereoselective catalytic cross-metathesis. *J. Am. Chem. Soc.* **135,** 6026–6029 (2013).
22. Bronner, S. M., Herbert, M. B., Patel, P. R., Marx, V. M. & Grubbs, R. H. Ruthenium-catalyzed *Z*-selective metathesis catalysts with modified cyclometalated carbene ligands. *Chem. Sci.* **5,** 4091–4098 (2014).
23. Brown, H. C., Larock, R. C., Gupta, S. K., Rajagopalan, S. & Bhat, N. G. Vinylic organoboranes. 15. Mercuration of 2-alkenyl-1,3,2-benzodioxaboroles and boronic acids. A convenient stereospecific procedure for the conversion of alkynes into (*E*)-1-halo-1-alkenes via mercuric salts. *J. Org. Chem.* **54,** 6079–6084 (1989).
24. Petasis, N. A. & Zavialov, I. A. Mild conversion of alkenyl boronic acids to alkenyl halides with halosuccinimides. *Tetrahedr. Lett.* **37,** 567–570 (1996).
25. Gensch, K. H., Pitman, I. H. & Higuchi, T. Oxidation of thioesters to sulfoxides by iodine. II. Catalytic role of some carboxylic acid anions. *J. Am. Chem. Soc.* **90,** 2096–2104 (1968).
26. Hamri, S., Rodríguez, J., Basset, J., Guillaumet, G. & Pujol, M. D. A convenient iodination of indoles and derivatives. *Tetrahedron* **68,** 6269–6275 (2012).
27. Speed, A. W. H., Mann, S. J., O'Brien, R. V., Schrock, R. R. & Hoveyda, A. H. Catalytic *Z*-selective cross-metathesis in complex molecule synthesis: a convergent stereoselective route to disorazole C$_1$. *J. Am. Chem. Soc.* **136,** 16136–16139 (2014).
28. Hanh, T. T. H., Hang, D. T. T., Minh, C. V. & Dat, N. T. Anti-inflammatory effects of fatty acids isolated from *Chromolaena odorata*. *Asian Pac. J. Trop. Med.* **4,** 760–763 (2011).
29. López, S., Fernández-Trillo, F., Midón, P., Castedo, L. & Saá, C. First stereoselective syntheses of (–)-siphonodiol and (–)-tetrahydrosiphonodiol, bioactive polyacetylenes from marine sponges. *J. Org. Chem.* **70,** 6346–6352 (2005).
30. Campbell, M. G. & Ritter, T. Late-stage fluorination: from fundamentals to application. *Org. Process Res. Dev.* **18,** 474–480 (2014).
31. Imaizumi, Y. *et al.* Molecular basis of pimarane compounds as novel activators of large-conductance Ca$^{2+}$-activated channel $\alpha$-subunit. *Mol. Pharmacol.* **62,** 836–846 (2002).
32. Macnaughtan, M. L., Johnson, M. J. A. & Kampf, J. W. Olefin metathesis reactions with vinyl halides: formation, observation, interception, and fate of the ruthenium–monohalomethylidene moiety. *J. Am. Chem. Soc.* **129,** 7708–7709 (2007).
33. Sashuk, V., Samojlowicz, C., Szadkowska, A. & Grela, K. Olefin cross-metathesis with vinyl halides. *Chem. Commun.* 2468–2470 (2008).
34. Macnaughtan, M. L., Gary, J. B., Gerlach, D. L., Johnson, M. J. A. & Kamf, J. W. Cross-metathesis of vinyl halides. Scope and limitations of ruthenium-based catalysts. *Organometallics* **28,** 2880–2887 (2009).
35. Meek, S. J., O'Brien, R. V., Llaveria, J., Schrock, R. R. & Hoveyda, A. H. Catalytic *Z*-selective olefin metathesis for natural product synthesis. *Nature* **471,** 461–466 (2011).
36. Vasiliu, M., Li, S., Arduengo, A. J. III & Dixon, D. A. Bond energies in models of the Schrock metathesis catalyst. *J. Phys. Chem. C* **115,** 12106–12120 (2011).
37. Rosebrugh, L. E., Herbert, M. B., Marx, V. M., Keitz, B. K. & Grubbs, R. H. Highly active ruthenium metathesis catalysts exhibiting unprecedented activity and *Z* selectivity. *J. Am. Chem. Soc.* **135,** 1276–1279 (2013).
38. Koh, M. J. *et al.* High-value alcohols and higher-oxidation-state compounds by catalytic *Z*-selective cross-metathesis. *Nature* **517,** 181–186 (2015).
39. Silverio, D. L. *et al.* Simple organic molecules as catalysts for enantioselective synthesis of amines and alcohols. *Nature* **494,** 216–221 (2013).
40. Hoye, R. C., Andersen, G. L., Brown, S. G. & Schultz, E. E. Total synthesis of clathculins A and B. *J. Org. Chem.* **75,** 7400–7403 (2010).
41. Jang, H., Zhugralin, A. R., Lee, Y. & Hoveyda, A. H. Highly selective methods for synthesis of internal ($\alpha$-) vinylboronates through efficient NHC–Cu-catalyzed hydroboration of terminal alkynes. Utility in synthesis and mechanistic basis for selectivity. *J. Am. Chem. Soc.* **133,** 7859–7871 (2011).
42. Hoveyda, A. H. Evolution of catalytic enantioselective olefin metathesis: from ancillary transformation to purveyor of stereochemical identity. *J. Org. Chem.* **79,** 4763–4792 (2014).
43. Fustero, S., Simón-Fuentes, A., Barrio, P. & Haufe, G. Olefin metathesis reactions with fluorinated substrates, catalysts, and solvents. *Chem. Rev.* **115,** 871–930 (2015).
44. Zhang, H., Yu, E. C., Torker, S., Schrock, R. R. & Hoveyda, A. H. Preparation of macrocyclic *Z*-enoates and (*E,Z*)- or (*Z,E*)-dienoates through catalytic stereoselective ring-closing metathesis. *J. Am. Chem. Soc.* **136,** 16493–16496 (2014).
45. Gillingham, D. G. & Hoveyda, A. H. Chiral N-heterocyclic carbenes in natural product synthesis: application of Ru-catalyzed asymmetric ring-opening/cross-metathesis and Cu-catalyzed allylic alkylation to total synthesis of baconipyrone C. *Angew. Chem. Int. Ed.* **46,** 3860–3864 (2007).
46. Addington, D. E. *et al.* Impact of second-generation anti-psychotics and perphenazine on depressive symptoms in a randomized trial for chronic schizophrenia. *J. Clin. Psychiatry* **72,** 75–80 (2011).
47. English, A. R., Girard, D., Jasys, V. J., Martingano, R. J. & Kellogg, M. S. Orally effective acid prodrugs of the $\beta$-lactamase inhibitor sulbactam. *J. Med. Chem.* **33,** 344–347 (1990).
48. Furuya, T. & Ritter, T. Fluorination of boronic acids mediated by silver(I) triflate. *Org. Lett.* **11,** 2860–2863 (2009).
49. Haginaka, J., Yasuda, H., Uno, T. & Nakagawa, T. Alkaline degradation and determination by high-performance liquid chromatography. *Chem. Pharm. Bull.* **32,** 2752–2758 (1984).
50. Townsend, E. M. *et al.* High oxidation state molybdenum imido heteroatom-substituted alkyliene complexes. *Organometallics* **32,** 4612–4617 (2013).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This research was supported by the United States National Institutes of Health, Institute of General Medical Sciences (GM-59426 and in part GM-57212).

**Author Contributions** M.J.K., T.T.N. and H.Z. were involved in the discovery, design and development of the new *Z*-selective cross-metathesis strategies and their applications. A.H.H., M.J.K., T.T.N. and H.Z. conceived the research programme. A.H.H. designed and directed the investigations. A.H.H. and R.R.S. conceived the studies that led to the development of Mo MAP complexes. A.H.H. wrote the manuscript with revisions provided by M.J.K., T.T.N. and H.Z.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.H.H. (amir.hoveyda@bc.edu).

doi:10.1038/nature17193

# Lytic to temperate switching of viral communities

B. Knowles[1]*, C. B. Silveira[1,2]*, B. A. Bailey[3], K. Barott[4], V. A. Cantu[5], A. G. Cobián-Güemes[1], F. H. Coutinho[2,6], E. A. Dinsdale[1,7], B. Felts[3], K. A. Furby[8], E. E. George[1], K. T. Green[1], G. B. Gregoracci[9], A. F. Haas[1], J. M. Haggerty[1], E. R. Hester[1], N. Hisakawa[1], L. W. Kelly[1], Y. W. Lim[1], M. Little[1], A. Luque[3,5,7], T. McDole-Somera[8], K. McNair[5], L. S. de Oliveira[2], S. D. Quistad[1], N. L. Robinett[1], E. Sala[10], P. Salamon[3,7], S. E. Sanchez[1], S. Sandin[8], G. G. Z. Silva[5], J. Smith[8], C. Sullivan[11], C. Thompson[2], M. J. A. Vermeij[12,13], M. Youle[14], C. Young[15], B. Zgliczynski[8], R. Brainard[15], R. A. Edwards[5,7,16], J. Nulton[3], F. Thompson[2] & F. Rohwer[1,7]

**Microbial viruses can control host abundances via density-dependent lytic predator–prey dynamics. Less clear is how temperate viruses, which coexist and replicate with their host, influence microbial communities. Here we show that virus-like particles are relatively less abundant at high host densities. This suggests suppressed lysis where established models predict lytic dynamics are favoured. Meta-analysis of published viral and microbial densities showed that this trend was widespread in diverse ecosystems ranging from soil to freshwater to human lungs. Experimental manipulations showed viral densities more consistent with temperate than lytic life cycles at increasing microbial abundance. An analysis of 24 coral reef viromes showed a relative increase in the abundance of hallmark genes encoded by temperate viruses with increased microbial abundance. Based on these four lines of evidence, we propose the Piggyback-the-Winner model wherein temperate dynamics become increasingly important in ecosystems with high microbial densities; thus 'more microbes, fewer viruses'.**

Microbial viruses infect about $10^{23}$ cells per second in the world's oceans and the majority of microbial cells are infected at any given time[1,2]. What determines the proportion of lytic versus lysogenic infections is not well understood, despite the known importance of lytic/lysogenic fate in driving ecological and biogeochemical outcomes[1,3–6]. Kill-the-Winner (KtW) models of lytic infection predict that density- and frequency-dependent viral predation suppresses blooms of rapidly growing hosts, increasing host diversity[6–9]. A number of studies provide empirical support for these predictions[7,10–13]. In contrast, temperate viral dynamics in the environment are much less studied, and the relationship between lysogeny and host density is unclear. Provirus induction studies indicate that lysogeny is more frequent with low host density[14–16]. As such, it was established that viral communities transition from lysogeny to lytic dominance as host densities rise[4,10,12,14,15,17].

Coral reefs offer a unique opportunity to probe the relationship between microbial host densities and the relative frequency of lytic versus temperate viral life cycles. Anthropogenic stressors can shunt these ecosystems into degradative regimes that result in changes in viral and microbial community composition, and rising microbial energy demand and densities, a state described as microbialized[18–23]. On heavily microbialized reefs, microbial abundances increase five- to tenfold, which increases predicted virus–host encounter rates[18,24]. Density-dependent lytic KtW models predict that reef microbialization should therefore correlate with increased lytic viral predation, resulting in a predicted increased virus to microbe ratio. Here we use four independent analyses—direct counts, literature meta-analyses, experiments, and viral community metagenomics—to show that increased host density is instead accompanied by a transition from lytic to temperate dynamics. On this basis we propose an extension of the KtW models, the Piggyback-the-Winner (PtW) model, which reflects the increased contribution of temperate viruses in ecosystems with high host abundance, yielding 'more microbes, fewer viruses'.

## Viral and microbial abundance

Microbial and viral abundances were measured in 223 Pacific and Atlantic coral reef samples (Fig. 1a). The density of virus-like particles (VLPs) was significantly higher than that of the microbes ($t = -19.61$, degrees of freedom (d.f.) $= 236.96$, $P < 2.20 \times 10^{-16}$; Welch two sample $t$-test) and ranged from $9.03 \times 10^5$ to $3.86 \times 10^7$ ($7.08 \times 10^6 \pm 3.01 \times 10^5$, mean ± standard error of the mean, s.e.m.) VLPs $\mathrm{ml}^{-1}$ versus $8.08 \times 10^4$ to $6.75 \times 10^6$ ($1.09 \times 10^6 \pm 5.53 \times 10^4$, mean ± s.e.m.) microbes $\mathrm{ml}^{-1}$. The log–log plot of these VLP and microbe abundances had a slope $<1$ ($m = 0.59$, $t = 14.82$, d.f. $= 221$, $P$ ($t$-test; $m \neq 1$) $= 4.08 \times 10^{-21}$; $R^2 = 0.50$; slope significantly different from $m = 1$ by linear regression with $t$-test; Fig. 1a), indicating a downward concave relationship between these variables. As a result, the virus to microbe ratio (VMR) decreased significantly (analysed against host density, both log-transformed; $m = -0.37$; $t = -9.52$, d.f. $= 221$, $P < 2.00 \times 10^{-16}$; $R^2 = 0.29$; linear regression) from a ratio of 25 to 2 VLPs per microbe ($7.44 \pm 0.24$, mean ± s.e.m.) as microbial abundance increased from $\sim 1 \times 10^5$ to greater than $6 \times 10^6$.

[1]Department of Biology, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA. [2]Biology Institute, Rio de Janeiro Federal University, Av. Carlos Chagas Filho 373, Rio de Janeiro, Rio de Janeiro 21941-599, Brazil. [3]Department of Mathematics and Statistics, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA. [4]Hawaii Institute of Marine Biology, University of Hawaii at Manoa, 46-007 Lilipuna Road, Kaneohe, Hawaii 96744, USA. [5]Computational Science Research Center, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA. [6]Radboud University Medical Centre, Radboud Institute for Molecular Life Sciences, Centre for Molecular and Biomolecular Informatics, 6525HP Nijmegen, The Netherlands. [7]Viral Information Institute, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA. [8]Scripps Institution of Oceanography, 8622 Kennel Way, La Jolla, California 92037, USA. [9]Marine Sciences Department, Sao Paulo Federal University - Baixada Santista, Av. Alm. Saldanha da Gama, 89, Santos, São Paulo 11030-400, Brazil. [10]National Geographic Society, 1145 17th St NW, Washington D.C. 20036, USA. [11]Department of Biology, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA. [12]CARMABI Foundation, Piscaderabaai z/n, Willemstad, Curacao, Netherlands Antilles. [13]Aquatic Microbiology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, 1098XH Amsterdam, The Netherlands. [14]Rainbow Rock, Ocean View, Hawaii 96737, USA. [15]Coral Reef Ecosystem Division-PIFSC-NOAA, 1845 Wasp Blvd, Honolulu, Hawaii 96818, USA. [16]Department of Computer Science, San Diego State University, 5500 Campanile Drive, San Diego, California 92182, USA.
*These authors contributed equally to this work.

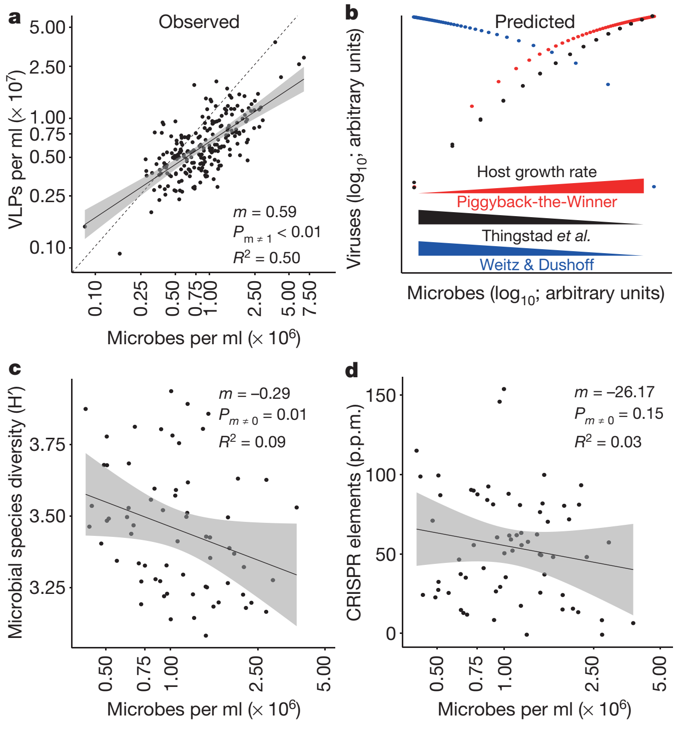


**Figure 1 | Virus-like particle (VLP) relative abundance declines with increasing host density despite lower microbial diversity and similar host sensitivity to infection, contrary to predictions of lytic models.** **a**, Log-transformed VLP versus microbial densities have an $m < 1$ relationship ($n = 223$ independent measures); the dashed reference line depicts a 10:1 relationship. **b**, Steady-state microbial and viral abundances and schematic microbial growth rate predicted by three modified Lotka–Volterra models: Piggyback-the-Winner (red), Thingstad *et al.* (2014; black[9]), and Weitz and Dushoff (2008; blue[8]). **c**, Shannon microbial species diversity versus host density (H′; $n = 66$ independent measures). **d**, Abundance of CRISPR elements in the microbial metagenomes ($n = 66$ independent measures). All slopes ($m$), $R^2$, and $P$ values describe linear regressions testing against a slope of 0, except **a** which shows the $P$ value from a two-sided $t$-test against a slope ≠ 1. Black best-fit lines with grey 99% prediction intervals from linear regressions are shown (**a**, **c**, and **d**).

Recent models were used to contrast our counts with predicted viral–host relationships[8,9]. Weitz and Dushoff (2008)[8], in which burst size is proportional to density-dependent microbial growth rate, predicts a negative relationship between viral and host density as viral predation causes declining host density with rising density-dependent host growth rate (Fig. 1b; details of steady state solution in Materials and Methods). The KtW-like model by Thingstad *et al.* (2014)[9], that incorporates terms for nested resistance to viral infection amongst multiple host strains[9,13], predicts an approximately downward concave relationship between viral and host abundances with the increasing dominance of slow growing, resistant hosts suppressing lytic dynamics as host density rises (Fig. 1b). The Piggyback-the-Winner (PtW) model introduced here predicts a relationship between VLP and host densities similar to Thingstad *et al.* (2014)[9], but lytic dynamics are suppressed at high host density and density-dependent growth rate owing to the increased prevalence of lysogeny (modelled as lower specific viral production rates per infection) and super-infection exclusion rather than resistance.

## Diversity and functional composition of microbial communities

Viral predation is thought to stimulate species-level host diversity through lineage-specific predation targeting dominant lineages, promoting community evenness[5,25]. However, when microbial diversity in 66 microbiomes from across the Pacific was probed, a weak and significantly negative relationship between host density and taxonomic diversity was observed (Fig. 1c; microbial abundance log-transformed; $m = -0.29$, $t = -2.60$, d.f. $= 64$, $P = 0.01$; $R^2 = 0.09$; linear regression). This also indicates that lytic dynamics are suppressed when both density-dependent (that is, total encounter rates) and frequency-dependent (that is, the relative density of a given host) would both favour lytic activity.

A recent model suggests that elevated host densities lead to an increase in host resistance to viral infection[9]. However, investigation of 66 Pacific microbiomes yielded weak relationships and no support for increased host resistance via CRISPRs or potential horizontal transfer of resistance (per cent competence genes) in the mixed microbial community metagenomes (CRISPRs: Fig. 1d; $m = -26.17$, $t = -1.44$, d.f. $= 64$, $P = 0.15$; $R^2 = 0.03$; per cent competence genes: Extended Data Fig. 1a; $m = -0.25$, $t = -2.40$, d.f. $= 64$, $P = 0.02$; $R^2 = 0.08$; microbial abundance log-transformed and linear regressions in both analyses). These data indicate that immunity to viral infection does not change with host density as predicted by Thingstad *et al.* (2014)[9], and is not promoted by horizontal transfer of resistance genes as predicted in King-of-the-Mountain dynamics[9,26]. Rather than host-mediated resistance to viral infection, the observed decrease in VMR with increasing microbial abundance may be driven by an alternative strain-level diversification mechanism, such as increasing resistance via lysogeny. Lysogeny, with its implicit super-infection immunity dynamic, would yield similar predictions to Thingstad *et al.* (2014)[9], albeit through a different mechanism, and could complement the nested infection design of the Thingstad *et al.* (2014)[9] model in future studies of resistance/growth trade-offs.

## Viral and host abundances in other ecosystems

Data from 22 independent studies were compiled for a meta-analysis to determine the generality of the 'more microbes, fewer viruses' observation. These studies spanned five orders of magnitude of microbial and VLP densities (Fig. 2; summary statistics in Extended Data Table 1; references in References for Methods). Analysis of log-transformed microbial and VLP abundances yielded slopes of significantly <1 in eight of the eleven environments. VMR therefore declined with increasing microbial density in disparate coastal and estuarine, coral reef, deep ocean, open ocean, temperate lake, animal-associated, sediment, and soil systems, consistent with our coral reef observations. This trend was also observed in the cystic fibrosis lung[27]. Together, these results show that 'more microbes, fewer viruses' is a common phenomenon. When viewed across the full range of host densities, peak VMR values were observed at ~$10^6$ microbes ml$^{-1}$ or g$^{-1}$ of sample. VMR declines as host density decreases or increases from this value (Fig. 2, final panel).

The relationship between microbial and viral densities was further examined through an analysis of published values of the fraction of lysogenic cells determined by mitomycin C induction[4,28–30]. Although a sometimes-significant negative relationship exists at a within-study level, examination across the full range of host abundances studied revealed no significant slope (Extended Data Fig. 2). The model that low host density favours lysogeny is not well supported by induction data when viewed globally; there is reason to re-examine the drivers of lysogeny with lines of evidence independent of established methods.

## Experimental manipulation of host growth rate

Our counts data contrast with predicted density-dependent lytic predation. Further, the models examined in Fig. 1b predict different relationships between microbial density, density-dependent host growth rate and viral lytic activity (measured as VMR). The actual relationship between these variables was probed with incubation experiments using seawater sampled from a pristine coral reef (Palmyra; 120-h time series) and a degraded embayment (Mission Bay; 72-h time series). Data were pooled within sites as high variability led to a lack of significant impact of dissolved organic carbon addition on host density ($t = 0.82$, d.f. $= 32.18$, $P = 0.42$; Welch two sample $t$-test with microbial

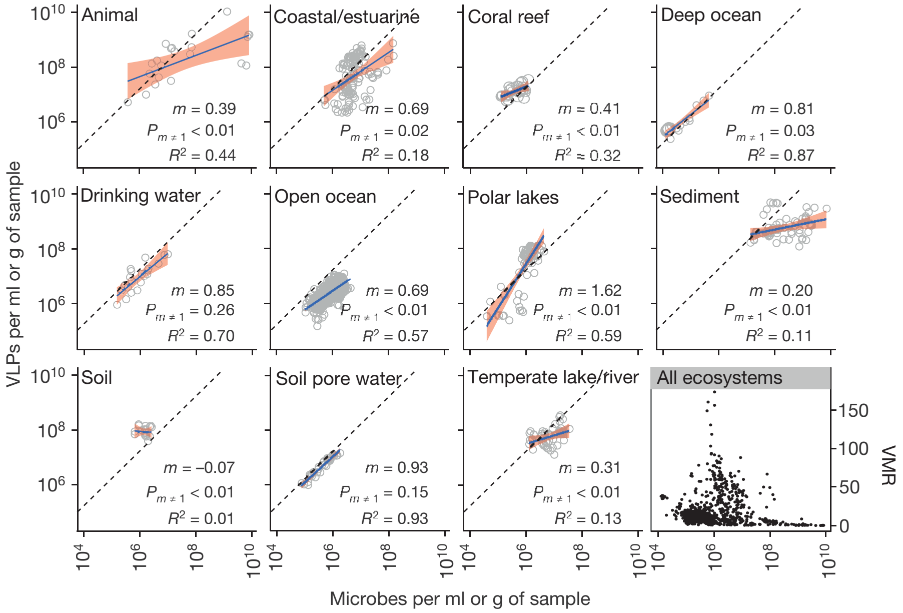


**Figure 2 | The relative decline in virus-like particles (VLPs) with increasing host density is common in disparate environmental systems.** Published microbial and VLP densities, and calculated virus to microbe ratio (with all environments pooled; final panel) are plotted by ecosystem. $n = 23$, 139, 27, 18, 22, 1397, 85, 71, 18, 35, and 46 independent measures for Animal-associated, Coastal/estuarine, Coral reef, Deep ocean, Drinking water, Open ocean, Polar lakes, Sediment, Soil, Soil pore water, and Temperate lake/river environments, respectively; pooled $n = 1{,}881$. Dashed lines depict 10:1 linear relationships; blue lines of best fit and pink 99% prediction intervals from linear regression are shown. All slopes ($m$) and $R^2$ values describe linear regressions, and $P$ values are derived from a two-sided $t$-test against a slope ≠ 1; details, including false-detection rate corrected values in Extended Data Table 1.

abundance log-transformed; Extended Data Fig. 3a) or VMR ($t = 0.17$, d.f. = 27.70, $P = 0.87$; Welch two sample $t$-test).

The experimental data matched our field observations: slopes significantly <1 were observed between log-transformed VLP and microbial densities in both incubations (Fig. 3a; Mission Bay: $m = 0.56$, $t = 6.65$, d.f. = 10, $P$ ($t$-test; $m \neq 1$) = $3.59 \times 10^{-4}$; $R^2 = 0.82$; Palmyra: $m = 0.63$, $t = 4.20$, d.f. = 23, $P$ ($t$-test; $m \neq 1$) = $2.25 \times 10^{-2}$; $R^2 = 0.43$; linear regression). These incubations are therefore more similar to the data set characterized by Wilcox and Fuhrman (1994)[3] as non-lytic (Fig. 3c; $m = 0.13$, $t = 0.64$, d.f. = 26, $P$ ($t$-test; $m \neq 1$) = $2.65 \times 10^{-4}$; $R^2 = 0.02$; linear regression) than the Hennes *et al.* (1995)[31] putatively lytic data set (Fig. 3c; $m = 1.19$, $t = 1.59$, d.f. = 4, $P$ ($t$-test; $m \neq 1$) = 0.81; $R^2 = 0.39$; linear regression). Hennes *et al.* (1995)[31] and Wilcox and Fuhrman (1994)[3] attribute their lytic (where VMR rose by ~40) and non-lytic dynamics to elevated and lowered microbial densities, respectively. In contrast, we did not observe a similar rise in VMR despite exceeding an order of magnitude higher host densities (Fig. 3b and 3d) and five times faster net growth rates than Hennes *et al.* (1995)[31] ($9.71 \times 10^6$ and $1.77 \times 10^6$ cells h$^{-1}$ in Palmyra and Mission Bay incubations, respectively; Hennes *et al.* (1995)[31]: $1.74 \times 10^6$ cells h$^{-1}$). These protist predator-free incubations (that is, 0.8 μm filtered) showed no significant increase in VMR with increasing host density, indicating that viral–host interactions alone are sufficient to drive the approximately downward concave relationship between VLP and host densities (Fig. 3a, b).

## Temperate genes, diversity, and virulence

Metagenomes of viral communities (viromes) from 24 Pacific and Atlantic coral reefs were sequenced (Extended Data Table 2). High variability and high leverage points were observed in the relationship of all viral bioinformatic metrics and host density, requiring the use of robust regression followed by bootstrap confidence interval estimation (RR-B) due to its insensitivity to high leverage, peripheral values. The per cent abundance of viral integrase, excisionase, and prophage reads increased significantly with microbial density (Fig. 4a, b, and Extended Data Fig. 4a) at the ≥90% confidence level (Fig. 4a, per cent integrase, $m = 1.23$, 90th percentile confidence interval (CI; 0.01, 2.69), Fig. 4b, per cent excisionase, $m = 0.04$, 90th CI (0.02, 0.10); Extended Data Fig. 4a, % prophage, $m = 0.13$, 90th CI (0.03, 0.44); RR-B against log-transformed host density; $R^2$ are not appropriate for robust regressions and are omitted).

Increased cell density was associated with a significant decline in functional diversity of the viral communities, an indicator of temperate viral communities[32], as measured by the Shannon (H′) index of putative coding genes in the viromes (Fig. 4c, $m = -3.54$, 90th CI (−6.14, −1.73); RR-B against log-transformed host density). Furthermore, the lower diversity, more temperate viral communities carry more virulence genes than the more diverse and lytic viral communities found at lower host densities (Fig. 4d, $m = 1.09$, 90th CI (0.46, 3.01); RR-B against log-transformed host density). Further, while we have conservatively used linear regression to analyse these relationships, the data suggests an exponential relationship between host density and % integrase, % excisionase, and % virulence genes, and a decay function between host density and viral functional diversity. These observed trends were not a result of overall viral community genome size reduction, as the average viral genome size determined by the Genome relative Abundance and Average Size tool was unaffected by microbial abundance ($m = -9190$, $t = -1.02$, d.f. = 22, $P = 0.32$; $R^2 = 0.04$; linear regression of genome length (bp) against log-transformed host density; mean estimated viral genome length = 42.07 ± 2.45 kb, mean ± s.e.m.). Rather, viral communities changed with increasing temperateness; viral communities clustered geographically as low-cell-density Atlantic viromes grouped away from Pacific viromes (Extended Data Fig. 4b).

## Discussion

The observed decline in the virus to microbe ratio (VMR) with elevated host abundances on microbialized coral reefs (Fig. 1) is consistent with lowered lytic activity at high host density (Fig. 1b). This trend was observed in eight of eleven (>70%) other disparate environments (Fig. 2). No support was found for competitive exclusion of viral predators by heterotrophic protists[6,33,34] (Fig. 3a), the rise or transfer of resistance to viral infection[9,26] (Fig. 1d and Extended Data Fig. 1b),

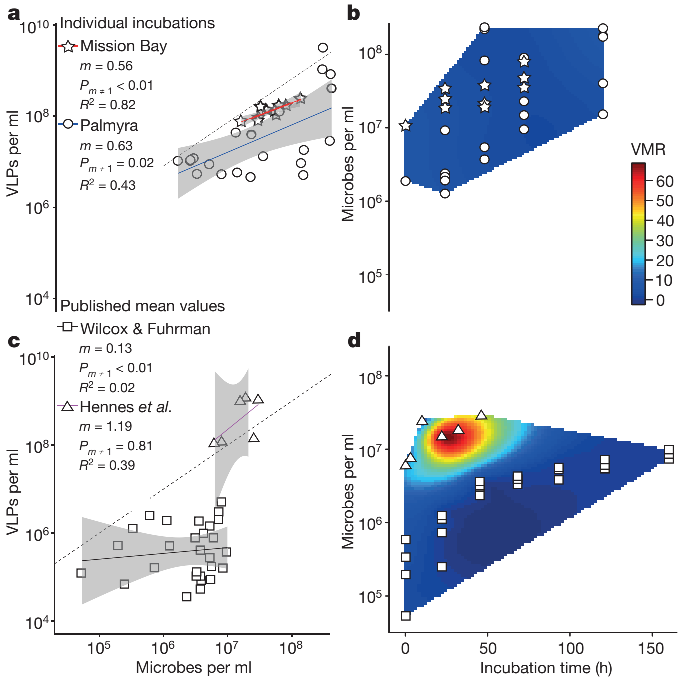


**Figure 3 | Density dependence does not drive viral predation. a**, **c**, Viral and host densities (individual counts shown) follow an $m < 1$ distribution despite high host densities (**a**; Mission Bay (stars) and Palmyra (circles) incubations; $n = 12$ and 25, respectively, from repeated measures), compared to putatively 'lytic' (slope $= 1$) and 'non-lytic' ($m < 1$) published data (**c**; Hennes *et al.* (1995)[31] (triangles) and Wilcox and Fuhrman (1994)[3] (squares), respectively; mean values shown; $n = 6$ and 28, respectively, from repeated measures). **b**, **d**, Microbe density and VMR over time in Mission Bay and Palmyra (**b**; individual values; $n = 3$ and 5 per time-point, respectively, except for time zero, when $n = 1$), and published putative lytic and non-lytic incubations (**d**; mean values; $n = 1$ and 4 per time-point, respectively) plotted over a thin plate spline. **a**, **c**, Dashed 10:1 lines, solid lines of best fit, with 99% prediction intervals in grey; all slopes ($m$) and $R^2$ values describe linear regressions, and $P$ values are derived from a two-sided $t$-test against a slope $\neq 1$. Individual incubation data are shown in Extended Data Fig. 3a. Mission Bay and Palmyra incubation experiments were each conducted once.

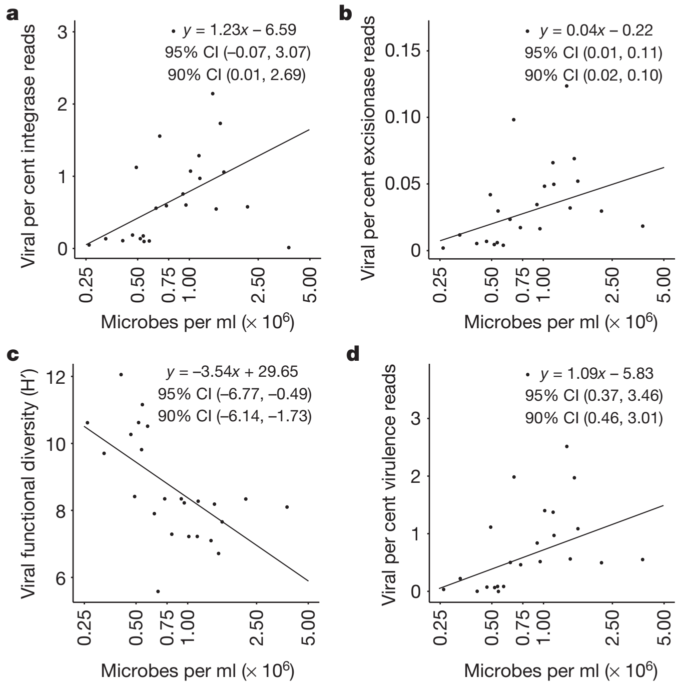


**Figure 4 | Temperate features in viromes increase with host density. a–d**, The relationship between log-transformed microbial density and the percent abundance of integrase (**a**), excisionase (**b**), and virulence reads in viromes (**d**), normalized by total sequences in each sample, and Shannon (H′) viral functional diversity (**c**) ($n = 24$ independent measures for all analyses). The linear equations and lines of best fit from robust regression and bootstrapped 95% and 90% confidence intervals (CIs) for the slopes are shown. Goodness of fit metrics are inappropriate for robust regression and are omitted.

greater species-level host diversity[5] (Fig. 1c), or increasing viral decay[5] (Extended Data Fig. 3b). Rather, multiple independent bioinformatic analyses of our viromes from this study, reinforced by viromes from other ecosystems[35–39], indicated an increased relative abundance of temperate viruses in communities with high microbial densities (Fig. 4). Empirical tests of alternative models to Piggyback-the-Winner (PtW) showed weak or ambiguous relationships while correlations supporting PtW were significant (for example, $R^2$ of $<0.09$ in Fig. 1c, d and Extended Data Fig. 2, compared with $R^2 > 0.56$ in Figs 1a and 3a). All four independent lines of evidence examined here—direct counts, literature meta-analyses, experiments, and viral community metagenomics—provide significant support for PtW.

The established model in viral ecology predicts that lytic dynamics dominate at high host density, whereas lysogeny is favoured at low densities[4,14,17]. We propose an extension of these Kill-the-Winner (KtW) models, Piggyback-the-Winner (PtW), wherein temperateness is favoured at high host densities as viruses exploit their hosts through lysogeny rather than killing them. As viral and host densities increase, lysogen resistance to superinfection by related viruses becomes increasingly important[9]. In this scenario, the energetic costs of generating resistance to infection through carrying proviruses should be less than through mutation[40,41]. Further, lysogeny can decouple microbial taxonomic and functional composition through horizontal gene transfer[42]. Virulence genes encode functions that harm eukaryotes; the increasing virulence content of viral communities under PtW dynamics (Fig. 4d) suggests that lysogens could evade protistan predation in addition to viral lysis. Suppressed top-down viral and protistan predation under PtW dynamics is likely to facilitate microbialization and ecosystem decline[22,23,43].

The 'narwhal-shaped' distribution that results when VMR is plotted against microbe density in multiple environments (Fig. 2, final panel) suggests that host densities observed in the ocean ($\sim 5 \times 10^5$ to $1 \times 10^6$ cells ml$^{-1}$) favour lytic KtW dynamics. Lower and higher host densities show a suppressed VMR. Thus, we predict that a Piggyback-the-Losers (PtL) dynamic extends the lytic-to-temperate shift to communities with low host densities. The diversity of environments across which the PtL–KtW–PtW dynamic is observed suggests that whichever viral–host dynamic prevails within a system, PtL, KtW or PtW, has major effects on processes as diverse as ecosystem function and disease progression[34,43–46].

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 19 May 2015; accepted 3 February 2016.**
**Published online 16 March 2016.**

1. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nature Rev. Microbiol.* **5,** 801–812 (2007).
2. Proctor, L. M. & Fuhrman, J. A. Viral mortality of marine bacteria and cyanobacteria. *Nature* **343,** 60–62 (1990).
3. Wilcox, R. M. & Fuhrman, J. A. Bacterial viruses in coastal seawater: lytic rather than lysogenic production. *Mar. Ecol. Ser.* **114,** 35–45 (1994).
4. Payet, J. & Suttle, C. A. To kill or not to kill: the balance between lytic and lysogenic viral infection is driven by trophic status. *Limnol. Oceanogr.* **58,** 465–474 (2013).
5. Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64,** 69–114 (2000).
6. Thingstad, T. F. Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnol. Oceanogr.* **45,** 1320–1328 (2000).

7. Rodriguez-Brito, B. *et al.* Viral and microbial community dynamics in four aquatic environments. *ISME J.* **4,** 739–751 (2010).
8. Weitz, J. S. & Dushoff, J. Alternative stable states in host–phage dynamics. *Theor. Ecol*. **1,** 13–19 (2008).
9. Thingstad, T. F., Våge, S., Storesund, J. E., Sandaa, R.-A. & Giske, J. A theoretical analysis of how strain-specific viruses can control microbial species diversity. *Proc. Natl Acad. Sci. USA* **111,** 7813–7818 (2014).
10. Weinbauer, M. G. & Höfle, M. G. Significance of viral lysis and flagellate grazing as factors controlling bacterioplankton production in a eutrophic lake. *Appl. Environ. Microbiol.* **64,** 431–438 (1998).
11. Bratbak, G., Egge, J. K. & Heldal, M. Viral mortality of the marine alga *Emiliania huxleyi* (Haptophyceae) and termination of algal blooms. *Mar. Ecol. Prog. Ser.* **93,** 39–48 (1993).
12. Evans, C. & Brussaard, C. P. D. Viral lysis and microzooplankton grazing of phytoplankton throughout the Southern Ocean. *Limnol. Oceanogr.* **57,** 1826–1837 (2012).
13. Needham, D. M. *et al.* Short-term observations of marine bacterial and viral communities: patterns, connections and resilience. *ISME J.* **7,** 1274–1285 (2013).
14. Jiang, S. C. & Paul, J. H. Significance of lysogeny in the marine environment: studies with isolates and a model of lysogenic phage production. *Microb. Ecol.* **35,** 235–243 (1998).
15. Paul, J. H. Prophages in marine bacteria: dangerous molecular time bombs or the key to survival in the seas? *ISME J.* **2,** 579–589 (2008).
16. Paul, J. H. & Weinbauer, M. Detection of lysogeny in marine environments. *Man. Aquat. Viral Ecol.* **4,** 30–33 (2010).
17. Maurice, C. F., Bouvier, C., de Wit, R. & Bouvier, T. Linking the lytic and lysogenic bacteriophage cycles to environmental conditions, host physiology and their variability in coastal lagoons. *Environ. Microbiol.* **15,** 2463–2475 (2013).
18. Dinsdale, E. A. *et al.* Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS One* **3,** e1584 (2008).
19. Smith, J. E. *et al.* Indirect effects of algae on coral: algae-mediated, microbe-induced coral mortality. *Ecol. Lett.* **9,** 835–845 (2006).
20. Thurber, R. V. *et al.* Metagenomic analysis of stressed coral holobionts. *Environ. Microbiol.* **11,** 2148–2163 (2009).
21. Kelly, L. W. *et al.* Black reefs: iron-induced phase shifts on coral reefs. *ISME J.* **6,** 638–649 (2012).
22. McDole, T. *et al.* Assessing coral reefs on a Pacific-wide scale using the microbialization score. *PLoS One* **7,** e43233 (2012).
23. Barott, K. L. & Rohwer, F. L. Unseen players shape benthic competition on coral reefs. *Trends Microbiol.* **20,** 621–628 (2012).
24. Alongi, D. M. *et al.* Phytoplankton, bacterioplankton and virioplankton structure and function across the southern Great Barrier Reef shelf. *J. Mar. Syst.* **142,** 25–39 (2015).
25. Thingstad, T. F. & Lignell, R. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat. Microb. Ecol.* **13,** 19–27 (1997).
26. Giovannoni, S., Temperton, B. & Zhao, Y. Giovannoni *et al.* reply. *Nature* **499,** E4–E5 (2013).
27. James, C. E. *et al.* Lytic activity by temperate phages of *Pseudomonas aeruginosa* in long-term cystic fibrosis chronic lung infections. *ISME J.* **9,** 1391–1398 (2015).
28. Muck, S. *et al.* Fracture zones in the Mid Atlantic Ridge lead to alterations in prokaryotic and viral parameters in deep-water masses. *Front. Microbiol*. **5,** 264 (2014).
29. Bongiorni, L., Magagnini, M., Armeni, M., Noble, R. & Danovaro, R. Viral production, decay rates, and life strategies along a trophic gradient in the North Adriatic Sea. *Appl. Environ. Microbiol.* **71,** 6644–6650 (2005).
30. Williamson, S. J., Houchin, L. A., McDaniel, L. & Paul, J. H. Seasonal variation in lysogeny as depicted by prophage induction in Tampa Bay, Florida. *Appl. Environ. Microbiol.* **68,** 4307–4314 (2002).
31. Hennes, K. P., Suttle, C. A. & Chan, A. M. Fluorescently labeled virus probes show that natural virus populations can control the structure of marine microbial communities. *Appl. Environ. Microbiol.* **61,** 3623–3627 (1995).
32. McDaniel, L. D., Rosario, K., Breitbart, M. & Paul, J. H. Comparative metagenomics: natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environ. Microbiol.* **16,** 570–585 (2014).
33. Fuhrman, J. A. Microbial community structure and its functional implications. *Nature* **459,** 193–199 (2009).
34. Weitz, J. S. *et al.* A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes. *ISME J.* **9,** 1352–1364 (2015).
35. Breitbart, M. *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J. Bacteriol.* **185,** 6220–6223 (2003).
36. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466,** 334–338 (2010).
37. Angly, F. E. *et al.* The marine viromes of four oceanic regions. *PLoS Biol.* **4,** e368 (2006).
38. Breitbart, M. *et al.* Diversity and population structure of a near-shore marine-sediment viral community. *Proc. R. Soc. Lond. B* **271,** 565–574 (2004).
39. Brum, J. R., Hurwitz, B. L., Schofield, O., Ducklow, H. W. & Sullivan, M. B. Seasonal time bombs: dominant temperate viruses affect Southern Ocean microbial dynamics. *ISME J.* **10,** 437–449 (2015).
40. Våge, S., Storesund, J. E. & Thingstad, T. F. Adding a cost of resistance description extends the ability of virus-host model to explain observed patterns in structure and function of pelagic microbial communities. *Environ. Microbiol.* **15,** 1842–1852 (2013).
41. Avrani, S. & Lindell, D. Convergent evolution toward an improved growth rate and a reduced resistance range in *Prochlorococcus* strains resistant to phage. *Proc. Natl Acad. Sci. USA* **112,** E2191–E2200 (2015).
42. Kelly, L. W. *et al.* Local genomic adaptation of coral reef-associated microbiomes to gradients of natural variability and anthropogenic stressors. *Proc. Natl Acad. Sci. USA* **111,** 10227–10232 (2014).
43. Silveira, C. B. *et al.* Microbial and sponge loops modify fish production in phase-shifting coral reefs. *Environ. Microbiol.* **17,** 3832–3846 (2015).
44. Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl Acad. Sci. USA* **110,** 10771–10776 (2013).
45. Whiteson, K. L. *et al.* The upper respiratory tract as a microbial source for pulmonary infections in cystic fibrosis. Parallels from island biogeography. *Am. J. Respir. Crit. Care Med.* **189,** 1309–1315 (2014).
46. Willner, D. *et al.* Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One* **4,** e7370 (2009).

**Acknowledgements** This paper is dedicated to the memory of Mike Furlan, mentor, friend, and colleague. We are grateful to the National Oceanographic and Atmospheric Administration Coral Reef Ecosystem Division for supporting this research, and to the captains and crews of the NOAA ship *Hi'ialakai* and the *Hanse Explorer*. Thanks to J. Payet for providing viral and microbial abundance data. Sampling was carried out under research permits from the US Fish and Wildlife Service, Palmyra Atoll National Wildlife Refuge, the Environment and Conservation Division of the Republic of Kiribati (n. 021/13) and ICMBio, Brazil (n. 27147-2). This work was funded by the Canadian Institute for Advanced Research Integrated Microbial Biodiversity Program Fellowship Award 141679 (to F.R.) and National Science Foundation grants OISE-1243541 and DEB-1046413 (to F.R.), CNS-1305112 and MCB-1330800 (to R.A.E.), DUE-1323809 (to E.A.D.), Gordon and Betty Moore Foundation Investigator Award GBMF-3781 (to F.R.), and the Brazilian National Research Council (CNPq; to F.T.) and Brazilian National Research Council Science Without Borders Program (CNPq/CAPES; to C.B.S.).

**Author Contributions** F.R., B.K., C.B.S., and F.T. conceptualized the project; B.K., F.R., C.B.S., and M.Y. wrote the manuscript; B.K., C.B.S., V.A.C., A.G.C.-G., K.T.G, K.M., G.G.Z.S., S.D.Q., Y.W.L., S.E.S., F.H.C., E.R.H. , N.L.R., B.A.B., B.F., A.L., P.S., J.N., C.Y., E.E.G., M.L., K.A.F., L.S.O., T.M.-S., J.M.H., B.Z., A.F.H., M.J.A.V., K.B., C.S., R.A.E., and F.R. performed sample collection, processing, experiments, and analysis; N.H. provided graphics and GIS analysis; E.A.D., L.W.K., S.S., J.S., R.B., C.T., G.B.G., J.N., E.S., R.A.E., F.T., and F.R. provided intellectual guidance and funding during the development of the research.

**Author Information** The viromes and microbiomes used in this paper are accessible at MG-RAST (http://metagenomics.anl.gov/) under the Piggyback-the-Winner project. Virome accession numbers: 4683670.3, 4683674.3, 4683677.3, 4683680.3, 4683683.3, 4683684.3, 4683686.3, 4683690.3, 4683702.3, 4683703.3, 4683704.3, 4683706.3, 4683712.3, 4683720.3, 4683739.3, 4683744.3, 4683745.3, 4683746.3, 4683747.3, 4683731.3, 4683733.3, 4683734.3, 4683718.3, 4684617.3. Microbiome accession numbers: 4683666.3, 4683667.3, 4683668.3, 4683669.3, 4683671.3, 4683672.3, 4683673.3, 4683675.3, 4683676.3, 4683678.3, 4683679.3, 4683681.3, 4683682.3, 4683685.3, 4683687.3, 4683688.3, 4683689.3, 4683691.3, 4683692.3, 4683693.3, 4683694.3, 4683695.3, 4683696.3, 4683697.3, 4683698.3, 4683699.3, 4683700.3, 4683701.3, 4683705.3, 4683707.3, 4683708.3, 4683709.3, 4683710.3, 4683711.3, 4683713.3, 4683714.3, 4683715.3, 4683716.3, 4683717.3, 4683719.3, 4683721.3, 4683722.3, 4683723.3, 4683724.3, 4683725.3, 4683726.3, 4683727.3, 4683728.3, 4683729.3, 4683732.3, 4683735.3, 4683736.3, 4683737.3, 4683738.3, 4683740.3, 4683741.3, 4683742.3, 4683743.3, 4683748.3, 4683749.3, 4683750.3, 4683751.3, 4683752.3, 4683753.3, 4683754.3, 4684616.3. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to F.R. (frohwer@gmail.com) or B.K. (benjaminwilliamknowles@gmail.com).

## METHODS

**Viral and microbial counts.** Seawater was collected in 2-l diver-deployed Niskin bottles at approximately 10 m depth within 30 cm of the benthos on coral reefs across the Pacific and Atlantic Oceans[47]. Samples were fixed with 2% final concentration paraformaldehyde within four hours of collection. Pacific Ocean samples were filtered and stained with SYBR Gold (Life Technologies, USA), mounted on slides and analysed by epifluorescence microscopy[47]. Atlantic Ocean samples were flash frozen and stored in liquid nitrogen until analysis on a BD FACSCalibur flow cytometer[48]. Investigators were blinded when conducting all counts in this study (environmental or experimental), with sites or incubation samples imaged and analysed in a random order and identified only after analysis.

**Predator–prey modelling.** Steady state solutions to the dynamic model of Weitz and Dushoff (2007)[8] were calculated under varying carrying capacities (K). The chemostat model of Thingstad *et al.* (2014)[9] was run for varying K, and the final point in the evolution of the system plotted.

A standard lytic model[49] that incorporates a logistic or trophic-state dependence for the microbe growth rate r is given by the equations

$$\delta N/\delta t = \mathrm{r} \bullet (1 - N/K) \bullet N - (d \bullet N) - (\phi \bullet N \bullet V)$$

$$\delta V/\delta t = (\beta \bullet \phi \bullet N/K \bullet N \bullet V) - (m \bullet V)$$

where $d$ and $m$ are, respectively, the trophic-independent death rates for microbes and phage, $N$ and $V$ are, respectively, microbial host and viral abundances, $\beta$ is the burst size, and $\phi$ is the adsorption coefficient. This corresponds to the Weitz–Dushoff model[8] with their parameter $a$ (the fractional reduction of lysis at carrying capacity term) set equal to 0.

In this case the specific viral production rate per microbe is given by the product $\beta \bullet \phi$. In the new PtW model of viral–host interactions proposed here we replace this product with the quantity $\beta \bullet \phi \bullet N/K$, suppressing viral production as the system moves away from $K$ (that is, $N/K$ becomes smaller) to simulate augmentation of lysogeny in eutrophic conditions. In this case $\beta \bullet \phi$ has the interpretation as the maximum value for the specific viral production rate per microbe. Steady state solutions of host and viral densities in the PtW model generated herein were calculated across a range of $K$ (Fig. 1b). All models are available as Matlab scripts from https://github.com/benjaminwilliamknowles/Piggyback-the-Winner.

**Meta-analysis of cell and viral abundances.** The relationships between published VLP and cell abundances from disparate environments were probed from 22 studies[17,28,44,50–68]. When abundances were not available, we used the WebPlotDigitizer tool to recover data from graphs (http://arohatgi.info/WebPlotDigitizer/app/). Samples were grouped by habitat: animal-associated, polar lakes, coastal/estuarine, coral reefs, deep ocean, drinking water, open ocean, sediment, soil, soil water and temperate lake/river. We similarly extracted data from published studies and tested the relationship between cell abundance and the frequency of lysogenic cells as studied by mitomycin C induction in previous studies from the Adriatic Basin, Arctic Shelf, Mid Atlantic Ridge and Tampa Bay[4,28–30].

**Metaviromic sampling and processing.** Viral metagenomic samples were collected at 24 reefs (Extended Data Table 2), a subset of sites sampled for counts as previously described[47]. Pacific viral concentrates were treated with 250 μl of chloroform per 50 ml of concentrate to destroy microbes and purified using CsCl step gradient ultracentrifugation[47]. Viral DNA was extracted using the formamide/phenol/chloroform isoamyl alcohol technique[47] and amplified using the Linker Amplified Sequencing Library approach[69] and sequenced on an Illumina MySeq platform (Illumina, USA). Atlantic viral concentrates were passed through a 0.22 μm filter and 250 μl of chloroform per 50 ml of concentrate was added to remove microbes, followed by ultracentrifugation for further concentration. DNA from Atlantic sites was extracted by the phenol/chloroform/isoamyl alcohol technique, amplified using multiple displacement amplification[20] and sequenced on an Ion Torrent sequencer (Life Sciences, USA). Microbial metagenomes were prepared by DNA extraction from the >0.22 μm fraction of the microbial community using Nucleospin Tissue Extraction kits (Macherey Nagel, Germany)[47] and sequencing on an Illumina MySeq platform (Illumina, USA).

**Bioinformatics.** Sequences less than 100 bp and with mean quality scores less than 25 were removed using PrinSeq[70]. Acceptable sequences were then dereplicated with TagCleaner[71] and potential contaminants matching lambda or human DNA sequences removed with DeconSeq[72]. Focusing on microbial reads, microbial metagenomes were taxonomically annotated based on *k*-mer similarity using FOCUS[73]. Rank-abundance tables were then used to calculate microbial species-level Shannon (base *e*) taxonomic diversity. For the virome analysis, protein sequences of all integrase, excisionase, and competence gene sequences on the NCBI RefSeq database (http://www.ncbi.nlm.nih.gov/refseq/) were downloaded and made into BLAST databases (makeblastdb command; BLAST version 2.2.29+, ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). The Virulence Factors of Pathogenic Bacteria Database (http://www.mgc.ac.cn/VFs/main.htm) was used as a protein database for virulence genes. The percentage of each sequence library composed of integrase, excisionase, competence, or virulence genes was computed as the number of sequences with >60 bp match at a 40% identity to database sequences identified using BLASTx, normalized by the total number of sequences in the virome. CRISPRs were identified in microbiomes using the CRISPR Recognition Tool (https://github.com/ajmazurie/CRT) and hits normalized to parts per million (p.p.m.) against total reads. The fraction of known prophage-like reads in the viromes, normalized by total sequences, was assessed by a stringent (e-value $10^{-10}$) BLAST against known prophages in cultured bacteria downloaded from NCBI (hosts (number of prophage): *Escherichia coli* (36), *Shigella flexneri* (31), *Salmonella enterica* (16), *Staphylococcus aureus* (14), *Xylella fastidiosa* (12), *Yersinia pseudotuberculosis* (11), *Yersinia pestis* (9), *Shewanella baltica* (8), *Streptococcus pyogenes* (7), *Pseudomonas syringae* (7), *Salmonella typhimurium* (6), *Xanthomonas campestris* (5), *Mycobacterium tuberculosis* (4), *Yersinia enterocolitica* (3), *Streptococcus agalactiae* (3), *Stenotrophomonas maltophilia* (3), *Pseudomonas putida* (3), *Staphylococcus haemolyticus* (2), *Streptomyces avermitilis* (1), *Streptococcus uberis* (1), *Listeria monocytogenes* (1), *Caulobacter sp.* (1)). For functional diversity analysis, reads of each virome were assembled using MIRA[74] followed by ORF calling using FragGenScan[75] and ORF clustering at 85% identity using CD-HIT[76] to build protein cluster databases. We then performed BLASTx of reads against clusters databases to assess the number of reads assigned to each protein cluster. An OTU-like table was built using each cluster as a rank unit and read counts as abundance. Shannon (base *e*) indexes were calculated using the VEGAN package in R (http://cran.r-project.org/web/packages/vegan/index.html). Average viral genome size estimates were performed using GAAS[77] and virome clustering was performed using crAss[78].

**Bioinformatic code availability.** The following codes and parameters were used for each step of the viral functional diversity analysis:

Assembly parameters used in Mira:

minimum overlap = 30 and minimum relative score = 90.

FragGeneScan code: ./run_FragGeneScan.pl -genome=[seq_file_name] -out=[output_file_name] -complete=0 -train=illumina_10.

CD-HIT code: cd-hit –i [input fastafilename.faa] -o [outputfilename]_85 -c 0.85 -n 5

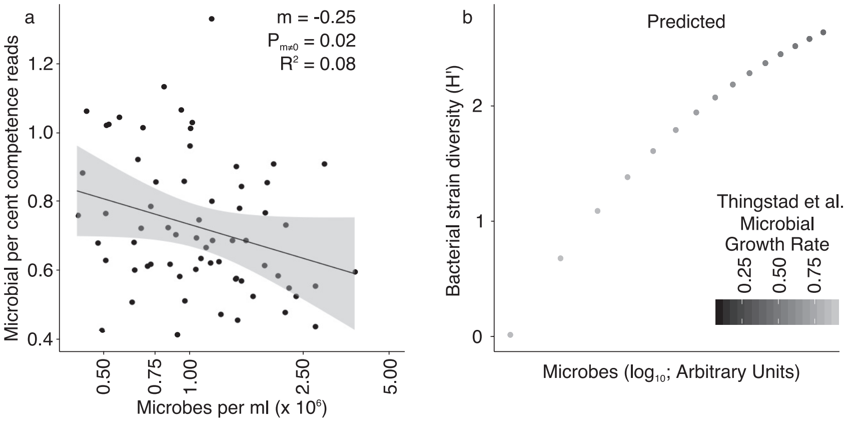
CD-HIT output was used as database for BLASTx with virome reads, and output format 6 was parsed with the following python script to create rank-abundance tables:

```
f="BlastOutput.txt"
myfile=open(f)
h={};temp=""
for line in myfile:
line=line.split()
if temp!=line[0]:
if line[1] not in h:
h[line[1]]=0
h[line[1]]+=1
temp=line[0]
myfile.close()
```

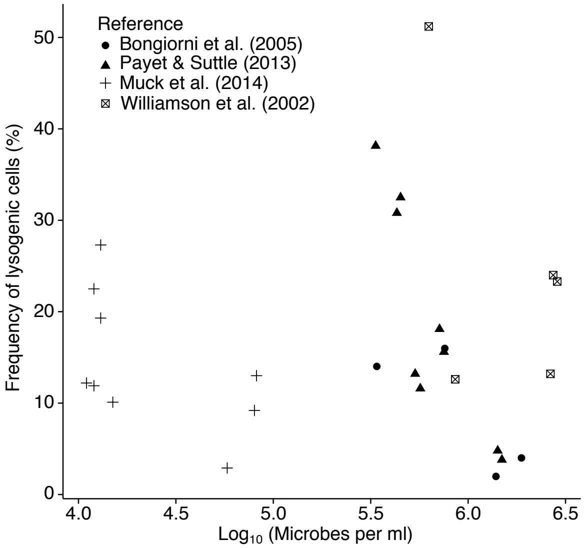
**Incubation experiments.** Water samples were collected at Palmyra Atoll, a pristine coral reef in the central Pacific, and Mission Bay, a degraded embayment in San Diego, CA. Samples were twice filtered through 0.8 μm pre-combusted GF/F filters to remove protists. Palmyra water was subsampled in 100-ml aliquots and distributed in 12 Whirl-Pak bags (Cole-Parmer, IL, USA), divided into two experimental groups and one control, each one containing four randomly chosen replicate bags. For the two experimental groups we added a DOC cocktail containing 48 different labile carbon sources[79] at the final concentration of 500 μM or 60 μM (+DOC treatment; Extended Data Fig. 3a), while no DOC was added to the control group (−DOC treatment; Extended Data Fig. 3a). Viral decay in microbe-free incubation bags was monitored as an additional control with 0.22 μm double-filtered water samples (Extended Data Fig. 3b). 1 ml samples were taken at times 0 h, 24 h, 48 h, 72 h, and 120 h from each bag for cell and viral counts. Mission Bay water was filtered and separated in three groups as above. 250 ml aliquots were distributed in each bag and incubated with 0 μM, 1 μM or 100 μM final concentrations by DOC addition. Samples were taken at times 12 h, 24 h, 48 h, and 72 h for counts. All incubations were performed in the dark at 25 °C. Samples were fixed and analysed by epifluorescence microscopy as described above.

**Statistical analysis.** No statistical methods were used to predetermine sample size. Significance was determined using an alpha of 0.05 when direct counts data were compared, and using an alpha of 0.1 when analysing counts versus bioinformatic analyses to account for the disparate nature of these data sets (although 95% prediction intervals are also shown). The relationship between microbial density and microbial diversity, CRISPR sequences, and competence genes were tested for significant deviation from a slope of 0 by linear regression. The relationship between VLP and microbial densities in Figs 1a, 2 (all except the final panel showing VMR), and 3a, c were tested for slopes significantly different to 1 by *t*-tests that tested the null hypothesis that the slope is not equal to 1 against the two-sided alternative; the corresponding *P* value is given for this test. The fdrtool package in R was used to provide false discovery rate-corrected (FDR) *P* values for the multiple comparisons conducted in Fig. 2 (Extended Data Table 1). Conclusions were similar between FDR and uncorrected analyses. Experimental data in Fig. 3 was complemented by average counts taken from previous studies[3,31] using the WebPlotDigitizer tool. Data was taken from the nutrient added treatment of Hennes *et al.* (1995)[31] as it was described as showing 'lytic' dynamics (Fig. 3c, d). Data from the 'non-lytic' 30%, 20%, 10%, and 3% dilutions by Wilcox and Fuhrman (1994)[3] were used as they had encounter rates (the product of viral and host densities) $\sim 10^{12}$ or less at the beginning of the incubations, described as the cutoff below which lytic dynamics were not sustained. A thin plate spline was applied to experimental and literature values for visualization and interpretation (Fig. 3b, d). While some data sets used to examine alternatives to PtW and published values in Figs 1c, d, 2, 3c, d, Extended Data Fig. 1a, and Extended Data Fig. 2, violated the assumptions of linear regression, this analysis was used for comparability. Robust regressions were used in Fig. 4 and Extended Data Fig. 4a analyses in order to accommodate high-leverage samples on parametric statistical models, allowing all samples to be retained in the analysis. Results are presented for robust regression estimation using Tukey's biweight and corresponding bootstrapped 90th percentile and 95th percentile confidence intervals (90% and 95% CIs) for the slope using 1,000 bootstrap replications. It should be noted for Fig. 4a that even though 95% CI covers 0, the 90% CI does not cover 0 indicating that there is evidence at the 0.1 confidence level that the slope is positive. For subsequent analyses in Fig. 4, 95% CIs do not straddle 0, showing that there is evidence at the 0.05 confidence level that the slope is negative (Fig. 4c) or positive (Fig. 4b, d). To account for error in the *y* axis we also performed Model II regression analyses with data shown in Figs 1, 2 and 4 using the package lmodel2 in R (Extended Data Table 3). It should be noted, however, that these results should be treated with caution, as error variance and goodness of fit metrics are not obtainable for this analysis.

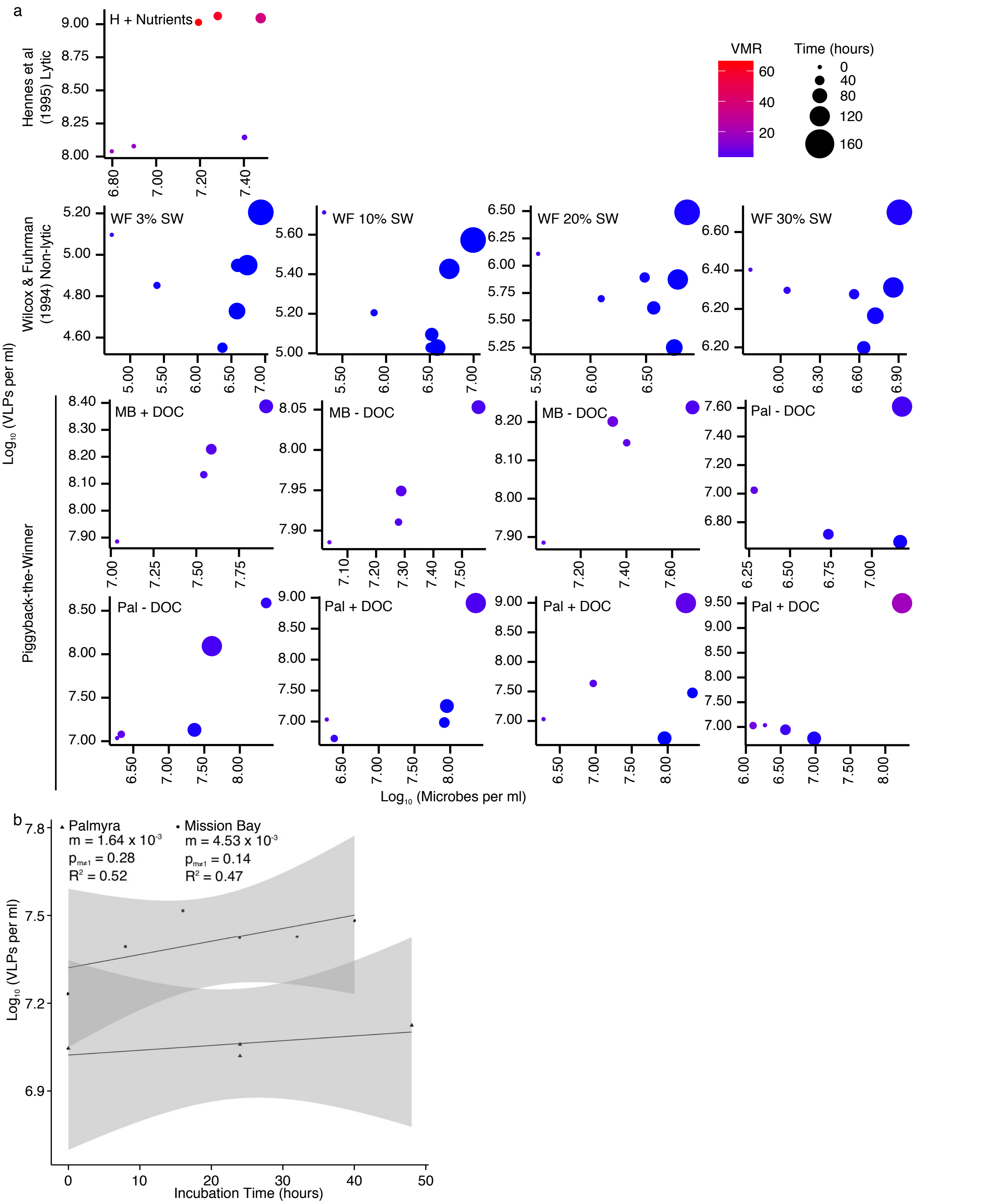
47. Haas, A. F. *et al.* Unraveling the unseen players in the ocean - a field guide to water chemistry and marine microbiology. *J. Vis. Exp.* **93,** e52131 (2014).
48. Brussaard, C. P. D., Payet, J. P., Winter, C. & Weinbauer, M. G. Quantification of aquatic viruses by flow cytometry. *Man. Aquat. Viral Ecol*. **11,** 102–109 (2010).
49. Murray, J. D. Mathematical biology: I. an introduction. (Springer, 2002).
50. Amossé, J. *et al.* The flows of nitrogen, bacteria and viruses from the soil to water compartments are influenced by earthworm activity and organic fertilization (compost vs. vermicompost). *Soil Biol. Biochem.* **66,** 197–203 (2013).
51. Bettarel, Y., Bouvy, M., Dumont, C. & Sime-Ngando, T. Virus-bacterium interactions in water and sediment of West African inland aquatic systems. *Appl. Environ. Microbiol.* **72,** 5274–5282 (2006).
52. Bouvier, T. & Maurice, C. F. A single-cell analysis of virioplankton adsorption, infection, and intracellular abundance in different bacterioplankton physiologic categories. *Microb. Ecol.* **62,** 669–678 (2011).
53. Glud, R. N. & Middelboe, M. Virus and bacteria dynamics of a coastal sediment: implication for benthic carbon cycling. *Limnol. Oceanogr.* **49,** 2073–2081 (2004).
54. Furlan, M. Viral and microbial dynamics in the human respiratory tract. (San Diego State Univ., 2009).
55. Hewson, I., O'Neil, J. M., Fuhrman, J. A. & Dennison, W. C. Virus-like particle distribution and abundance in sediments and overlying waters along eutrophication gradients in two subtropical estuaries. *Limnol. Oceanogr.* **46,** 1734–1746 (2001).
56. Kim, M.-S., Park, E.-J., Roh, S. W. & Bae, J.-W. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl. Environ. Microbiol.* **77,** 8062–8070 (2011).
57. Lisle, J. T. & Priscu, J. C. The occurrence of lysogenic bacteria and microbial aggregates in the lakes of the McMurdo Dry Valleys, Antarctica. *Microb. Ecol.* **47,** 427–439 (2004).
58. Laybourn-Parry, J., Marshall, W. A. & Madan, N. J. Viral dynamics and patterns of lysogeny in saline Antarctic lakes. *Polar Biol.* **30,** 351–358 (2006).
59. Madan, N. J., Marshall, W. a. & Laybourn-Parry, J. Virus and microbial loop dynamics over an annual cycle in three contrasting Antarctic lakes. *Freshw. Biol.* **50,** 1291–1300 (2005).
60. Maurice, C. F., Bouvier, T., Comte, J., Guillemette, F. & Del Giorgio, P. A. Seasonal variations of phage life strategies and bacterial physiological states in three northern temperate lakes. *Environ. Microbiol.* **12,** 628–641 (2010).
61. Maurice, C. F. *et al.* Disentangling the relative influence of bacterioplankton phylogeny and metabolism on lysogeny in reservoirs and lagoons. *ISME J.* **5,** 831–842 (2011).
62. Mei, M. L. & Danovaro, R. Virus production and life strategies in aquatic sediments. *Limnol. Oceanogr.* **49,** 459–470 (2004).
63. Parsons, R. J., Breitbart, M., Lomas, M. W. & Carlson, C. A. Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *ISME J.* **6,** 273–284 (2012).
64. Parsons, R. J. *et al.* Marine bacterioplankton community turnover within seasonally hypoxic waters of a subtropical sound: Devil's Hole, Bermuda. *Environ. Microbiol.* **17,** 3481–3499 (2015).
65. Rinta-Kanto, J. M., Lehtola, M. J., Vartiainen, T. & Martikainen, P. J. Rapid enumeration of virus-like particles in drinking water samples using SYBR green I-staining. *Water Res.* **38,** 2614–2618 (2004).
66. Schapira, M. *et al.* Distribution of heterotrophic bacteria and virus-like particles along a salinity gradient in a hypersaline coastal lagoon. *Aquat. Microb. Ecol.* **54,** 171–183 (2009).
67. Payet, J. P., McMinds, R., Burkepile, D. E. & Vega Thurber, R. L. Unprecedented evidence for high viral abundance and lytic activity in coral reef waters of the South Pacific Ocean. *Front. Microbiol*. **5,** 493 (2014).
68. Patten, N. L., Harrison, P. L. & Mitchell, J. G. Prevalence of virus-like particles within a staghorn scleractinian coral (*Acropora muricata*) from the Great Barrier Reef. *Coral Reefs* **27,** 569–580 (2008).
69. Duhaime, M. B., Deng, L., Poulos, B. T. & Sullivan, M. B. Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ. Microbiol.* **14,** 2526–2537 (2012).
70. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).
71. Schmieder, R., Lim, Y. W., Rohwer, F. & Edwards, R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* **11,** 341 (2010).
72. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* **6,** e17288 (2011).
73. Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E. & Edwards, R. A. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* **2,** e425 (2014).
74. Chevreux, B., Wetter, T. & Suhai, S. Genome sequence assembly using trace signals and additional sequence information. in *German conference on bioinformatics* 45–56 (1999).
75. Rho, M., Tang, H. & Ye, Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* **38,** e191 (2010).
76. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).
77. Angly, F. E. *et al.* The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLOS Comput. Biol.* **5,** e1000593 (2009).
78. Dutilh, B. E. *et al.* Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* **28,** 3225–3231 (2012).
79. Sanchez, S. E. *et al.* Phage Phenomics: physiological approaches to characterize novel viral proteins. *J. Vis. Exp.* **100,** e52854 (2015).



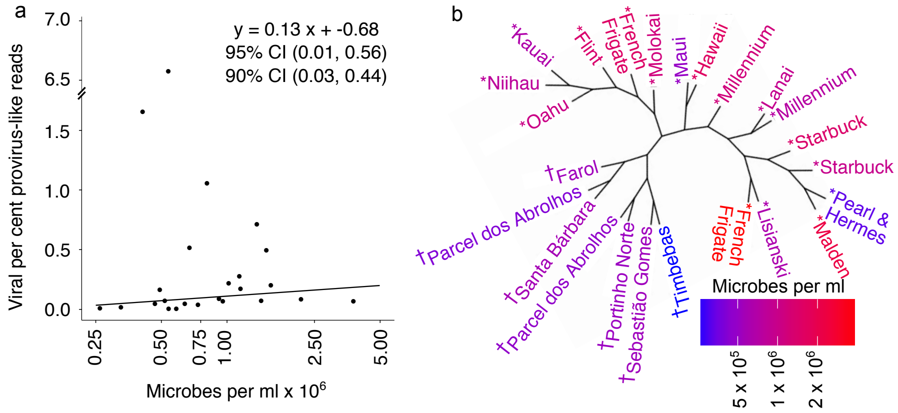
**Extended Data Figure 1 | The observed decline in virus to microbe ratio with increasing host density is not supported by horizontal transfer (for example, of resistance genes) under conditions where strain diversity is predicted to rise. a**, Host competence gene composition likely does not facilitate the expected rise in resistance to viral infection ($n = 66$; $m = -0.25$, $t = -2.40$, d.f. = 64, $P = 0.02$; $R^2 = 0.08$; microbial abundance log-transformed; linear regression). **b**, Lysogeny may provide strain diversification similar to the co-evolutionary diversification predicted by Thingstad *et al.* (2014)[9] nested-infection chemostat model.



**Extended Data Figure 2 | Meta-analysis of the frequency of lysogenic cells (FLC) from mitomycin C induction experiments yields ambiguous results.** FLC from four published studies is plotted against total cell abundance. Although a sometimes-significant negative relationship exists at a within-study level (microbial abundance log-transformed; Muck *et al.* (2014)[28], $n = 9$, $m = -10.79$, $t = -1.76$, d.f. = 7, $P = 0.12$; $R^2 = 0.31$; Bongiorni *et al.* (2005)[29], $n = 4$, $m = -17.23$, $t = -1.91$, d.f. = 2, $P = 0.20$; $R^2 = 0.65$; Payet and Suttle (2013)[4], $n = 9$, $m = -48.31$, $t = -4.80$, d.f. = 7, $P = 1.96 \times 10^{-3}$; $R^2 = 0.77$; Williamson *et al.* (2002)[30], $n = 5$, $m = -26.08$, $t = -1.08$, d.f. = 3, $P = 0.36$; $R^2 = 0.28$; linear regression of each data set examined independently), when examined altogether across the full range of host abundances studied, no significant slope was observed (microbial abundance log-transformed; $n = 27$, $m = -0.11$, $t = -0.04$, d.f. = 25, $P = 0.97$; $R^2 = 5.94 \times 10^{-5}$; linear regression of pooled data).



**Extended Data Figure 3 | Decline in virus to microbe ratio (VMR) observed in incubations with elevated host density over time, contrasted with published values and viral decay. a**, Log-transformed VLP density in experimental incubations is plotted against microbial host density over time (dot size) with VMR indicated by dot colour. Data from Mission Bay (MB) and Palmyra (Pal) water with DOC added (+ DOC) or not (− DOC) is complemented by the nutrient-added 'lytic' system of Hennes *et al.* (1995)[31] (H + Nutrients) as well as the 'non-lytic' dilutions (3%, 10%, 20%, and 30% final concentration seawater diluted by 0.02 μm filtered seawater) of Wilcox and Fuhrman (1994)[3]; WF 3% SW, WF 10% SW, WF 20% SW, WF 30% SW). $n = 1$ all incubations and published mean values. **b**, Significant viral decay was not observed in cell-free viral decay controls in incubation experiments (Palmyra: $n = 4$, $m = 1.64 \times 10^{-3}$, $t = 1.48$, d.f. = 2, $P = 0.28$; $R^2 = 0.52$; Mission Bay: $n = 6$, $m = 4.53 \times 10^{-3}$, $t = 1.87$, d.f. = 4, $P = 0.14$; $R^2 = 0.47$; linear regression with log-transformed viral density).



**Extended Data Figure 4 | Temperateness of viral communities increases with host density and viral functional composition change. a**, The relative composition of provirus-like reads, normalized by total sequences in each sample, increases with host density in viral metagenomes (host density log-transformed; $n = 24$ independent measures). The linear equations and line of best fit from robust regression and bootstrapped 95% and 90% confidence intervals (CIs) for the slope are shown. Goodness of fit metrics are inappropriate for robust regression and are omitted. **b**, Viromes clustered by functional similarity (crAss cross-assembly), showing higher host density Pacific viromes (*) grouped away from lower host density Atlantic viromes (†); site names coloured by host density.

**Extended Data Table 1 | Summary of linear regression analyses of published microbial and viral counts**

| Ecosystem | Intercept | Slope | $R^2$ | p-value (m≠1) | FDR |
|---|---|---|---|---|---|
| Animal | 5.34366 | 0.38784 | 0.4374 | **0.00000** | **0.00000** |
| Coastal/Estuarine | 3.03460 | 0.68960 | 0.1780 | **0.01548** | **0.02123** |
| Coral Reef | 4.84280 | 0.40500 | 0.3197 | **0.00003** | **0.00007** |
| Deep Ocean | 2.25335 | 0.80947 | 0.8713 | **0.02618** | **0.03200** |
| Drinking Water | 1.84750 | 0.8549 | 0.6967 | 0.26342 | 0.26342 |
| Open Ocean | 2.63034 | 0.68802 | 0.5708 | **0.00000** | **0.00000** |
| Polar Lake | -2.2782 | 1.6215 | 0.5933 | **0.00006** | **0.00011** |
| Sediment | 7.07985 | 0.20243 | 0.1062 | **0.00000** | **0.00000** |
| Soil | 8.40482 | -0.07364 | 0.0059 | **0.00037** | **0.00059** |
| Soil Pore Water | 1.42619 | 0.93485 | 0.9328 | 0.14518 | 0.15970 |
| Temperate Lake and River | 5.75600 | 0.3124 | 0.1341 | **0.00000** | **0.00000** |

Data shown in Fig. 2. Slope, intercept and $R^2$ are reported for each ecosystem, followed by the *P* value for two-tailed *t*-tests testing for slopes different from 1 and false discovery rates (FDR). Significant values are highlighted in bold. FDR-correction yielded similar results to uncorrected linear regressions.

**Extended Data Table 2 | Summary information on the post-quality control viromes analysed**

| Site | Reef area | Year | Sequencing Technology | Size (reads) | Size (bp) |
|---|---|---|---|---|---|
| Millenium | Southern Line Islands | 2013 | Illumina | 396,009 | 108,504,986 |
| French Frigate | Northwestern Hawaiian Islands | 2013 | Illumina | 27,811 | 7,620,127 |
| French Frigate | Northwestern Hawaiian Islands | 2013 | Illumina | 341,762 | 93,642,528 |
| Flint | Southern Line Islands | 2013 | Illumina | 26,649 | 7,288,052 |
| Hawaii | Main Hawaiian Islands | 2013 | Illumina | 169,145 | 46,345,548 |
| Kauai | Main Hawaiian Islands | 2013 | Illumina | 53,601 | 14,659,196 |
| Lanai | Main Hawaiian Islands | 2013 | Illumina | 216,678 | 59,369,605 |
| Lilianski | Northwestern Hawaiian Islands | 2013 | Illumina | 197,754 | 54,184,414 |
| Maui | Main Hawaiian Islands | 2013 | Illumina | 210,548 | 57,689,978 |
| Malden | Southern Line Islands | 2013 | Illumina | 273,248 | 74,869,874 |
| Millennium | Southern Line Islands | 2013 | Illumina | 205,734 | 56,370,772 |
| Molokai | Main Hawaiian Islands | 2013 | Illumina | 131,338 | 35,986,536 |
| Niihau | Main Hawaiian Islands | 2013 | Illumina | 62,462 | 17,086,857 |
| Oahu | Main Hawaiian Islands | 2013 | Illumina | 104,948 | 28,702,185 |
| Pearl & Hermes | Northwestern Hawaiian Islands | 2013 | Illumina | 896,672 | 245,686,926 |
| Starbuck | Southern Line Islands | 2013 | Illumina | 209,210 | 57,322,864 |
| Starbuck | Southern Line Islands | 2013 | Illumina | 158,006 | 43,293,430 |
| Farol | Abrolhos | 2011 | Ion Torrent | 5,309,048 | 1,111,783,405 |
| Parcel dos Abrolhos | Abrolhos | 2011 | Ion Torrent | 3,345,804 | 656,547,520 |
| Parcel dos Abrolhos | Abrolhos | 2012 | Ion Torrent | 870,631 | 182,798,753 |
| Portinho Norte | Abrolhos | 2012 | Ion Torrent | 718,794 | 143,135,114 |
| Sebastião Gomes | Abrolhos | 2012 | Ion Torrent | 317,401 | 67,968,285 |
| Santa Barbara | Abrolhos | 2011 | Ion Torrent | 323,883 | 68,006,354 |
| Timbebas | Abrolhos | 2012 | Ion Torrent | 933,142 | 149,255,589 |

Site, region sampled, the year samples were taken, sequencing platform, and the number of reads and base pairs (bp) in each virome are shown.

**Extended Data Table 3 | Summary of model II OLS, MA, and SMA regression analyses**

| Variables tested | Method | Intercept | Slope | Slope Confidence Intervals (2.5 % / 97.5 %) | P-perm |
|---|---|---|---|---|---|
| Log10(VLP abundance) ~ | OLS | 3.270594 | 0.5909753 | 0.5123641/0.6695865 | 0.01 |
| Log10(Microbial abundance) | MA | 2.154908 | 0.7784587 | 0.6796619/0.8877203 | 0.01 |
| Figure 1a | SMA | 1.805399 | 0.8371912 | 0.7622627/0.9194850 | NA |
| Microbial Species Diversity (H') ~ | OLS | 5.201590 | -0.2899249 | -0.5162249/-0.06362489 | 0.01 |
| Log10(Microbial abundance) | MA | 8.577445 | -0.8499407 | -1.9640454/-0.31929823 | 0.01 |
| Figure 1c | SMA | 9.189515 | -0.9514762 | -1.2043178/-0.75171774 | NA |
| CRISPR elements (ppm) ~ | OLS | 212.9163 | -26.1718 | -62.41016/10.06656 | 0.07 |
| Log10(Microbial abundance) | MA | 5063.2712 | -830.7903 | 2159.95654/-348.39257 | 0.07 |
| Figure 1d | SMA | 944.0547 | -147.4593 | -188.08521/-115.60849 | NA |
| Competence gene abundance | OLS | 2.224755 | -0.2484529 | -0.455530/-0.04137585 | 0.01 |
| ~ Log10(Microbial abundance) | MA | 4.438175 | -0.6156341 | -1.349562/-0.17225329 | 0.01 |
| Extended Data Figure 1a | SMA | 5.945424 | -0.8656695 | -1.097170/-0.68301541 | NA |
| Log10(VLP abundance) ~ | OLS | 5.343661 | 0.3878362 | 0.1882105/0.5874618 | 0.01 |
| Log10(Microbial abundance) | MA | 4.767191 | 0.4639165 | 0.2410987/0.7324118 | 0.01 |
| Figure 2h - Animal | SMA | 3.838768 | 0.5864464 | 0.4198659/0.8191172 | NA |
| Log10(VLP abundance) ~ | OLS | 3.034594 | 0.6896393 | 0.4393031/0.9399756 | 0.01 |
| Log10(Microbial abundance) | MA | -11.009437 | 2.7827455 | 2.0144638/4.2863025 | 0.01 |
| Figure 2a – Coastal/estuarine | SMA | -3.304430 | 1.6344001 | 1.4031244/1.9037969 | NA |
| Log10(VLP abundance) ~ | OLS | 4.842831 | 0.4049663 | 0.1616585/0.6482740 | 0.02 |
| Log10(Microbial abundance) | MA | 3.943664 | 0.5655113 | 0.2599340/0.9794991 | 0.02 |
| Figure 2b – Coral reefs | SMA | 3.099839 | 0.7161750 | 0.5130687/0.9996842 | NA |
| Log10(VLP abundance) ~ | OLS | 2.253353 | 0.8094707 | 0.6445947/0.9743466 | 0.01 |
| Log10(Microbial abundance) | MA | 2.031353 | 0.8584923 | 0.6967851/1.0509648 | 0.01 |
| Figure 2c- Deep ocean | SMA | 1.991944 | 0.8671945 | 0.7178530/1.0476048 | NA |
| Log10(VLP abundance) ~ | OLS | 1.8475372 | 0.8548761 | 0.5918003/1.117952 | 0.01 |
| Log10(Microbial abundance) | MA | 0.8154567 | 1.0290240 | 0.7527892/1.410628 | 0.01 |
| Figure 2f – Drinking water | SMA | 0.8442242 | 1.0241699 | 0.7943422/1.320494 | NA |
| Log10(VLP abundance) ~ | OLS | 2.630340 | 0.6880182 | 0.6566834/0.7193530 | 0.01 |
| Log10(Microbial abundance) | MA | 1.560510 | 0.8836194 | 0.8441542/0.9246933 | 0.01 |
| Figure 2d – Open ocean | SMA | 1.412589 | 0.9106645 | 0.8798687/0.9425383 | NA |
| Log10(VLP abundance) ~ | OLS | -2.278235 | 1.621498 | 1.328398/1.914598 | 0.01 |
| Log10(Microbial abundance) | MA | -7.623837 | 2.514119 | 2.120363/3.053842 | 0.01 |
| Figure 2e – Antarctic lakes | SMA | -5.174698 | 2.105156 | 1.832362/2.418562 | NA |
| Log10(VLP abundance) ~ | OLS | 7.079846 | 0.2024254 | 0.06135500/0.3434959 | 0.01 |
| Log10(Microbial abundance) | MA | 6.275650 | 0.3000129 | 0.09877745/0.5263175 | 0.01 |
| Figure 2i - Sediment | SMA | 3.628034 | 0.6212953 | 0.49603921/0.7781801 | NA |
| Log10(VLP abundance) ~ | OLS | 8.404815 | -0.07364098 | -0.581317 0.4340350 | 0.45 |
| Log10(Microbial abundance) | MA | 11.689048 | -0.60574792 | NA/NA | 0.45 |
| Figure 2i - Soil | SMA | 13.880157 | -0.96074842 | -1.594310/-0.5789576 | NA |
| Log10(VLP abundance) ~ | OLS | 1.426192 | 0.9348495 | 0.8460122/1.023687 | 0.01 |
| Log10(Microbial abundance) | MA | 1.246901 | 0.9668013 | 0.8788657/1.063210 | 0.01 |
| Figure 2g – Soil pore water | SMA | 1.240642 | 0.9679167 | 0.8831477/1.060822 | NA |
| Log10(VLP abundance) ~ | OLS | 5.756037 | 0.3124330 | 0.07122673/0.5536392 | 0.02 |
| Log10(Microbial abundance) | MA | 3.565635 | 0.6552208 | 0.22182386/1.3743755 | 0.02 |
| Figure 2h – Temperate lakes | SMA | 2.300832 | 0.8531567 | 0.64539226/1.1278047 | NA |
| Integrase abundance ~ | OLS | -4.124376 | 0.8290249 | -0.3506466/2.008696 | 0.07 |
| Log10(Microbial abundance) | MA | -48.478845 | 8.3292824 | 3.3669820/-20.091584 | 0.07 |
| Figure 4a | SMA | -15.743870 | 2.7938596 | 1.8530297/4.212373 | NA |
| Excisionase abundance ~ | OLS | -0.2354934 | 0.04659967 | -0.01650558/0.1097049 | 0.07 |
| Log10(Microbial abundance) | MA | -0.2412107 | 0.04756645 | -0.01678433/0.1123127 | 0.07 |
| Figure 4b | SMA | -0.8477914 | 0.15013811 | 0.09975581/0.2259663 | NA |
| Provirus-like sequences ~ | OLS | 5.291734 | -0.8009354 | -3.159957/1.558086 | 0.17 |
| Log10(Microbial abundance) | MA | 208.253969 | -35.1214691 | 18.023316/-8.862742 | 0.17 |
| Figure 4c | SMA | 32.460430 | -5.3951110 | -8.247333/-3.529289 | NA |
| Genetic diversity (H') ~ | OLS | 29.31875 | -3.492783 | -5.583645/-1.401920 | 0.01 |
| Log10(Microbial abundance) | MA | 66.09613 | -9.711770 | -24.038458/-6.056465 | 0.01 |
| Figure 4d | SMA | 43.42959 | -5.878899 | -8.330507/-4.148782 | NA |
| Pathogenicity gene abundance ~ | OLS | -6.065022 | 1.172082 | -0.2624953/2.60666 | 0.06 |
| Log10(Microbial abundance) | MA | -54.762269 | 9.406695 | 4.1780142/-43.26440 | 0.06 |
| Figure 5e | SMA | -19.534564 | 3.449756 | 2.3015745/5.17073 | NA |

The first column indicates the variables tested and the corresponding figures in the main text. Slopes, intercepts, confidence intervals and *P* values are shown. Rows with confidence intervals not covering 1 for Figs 1a and 2, or not covering 0 for Figs 1c, d, and 4, are significant.

doi:10.1038/nature17157

# Deletions linked to *TP53* loss drive cancer through p53-independent mechanisms

Yu Liu[1,2]*, Chong Chen[1,2]*, Zhengmin Xu[1], Claudio Scuoppo[3], Cory D. Rillahan[2], Jianjiong Gao[4], Barbara Spitzer[5,6], Benedikt Bosbach[2], Edward R. Kastenhuber[2], Timour Baslan[2], Sarah Ackermann[2], Lihua Cheng[7], Qingguo Wang[4], Ting Niu[7], Nikolaus Schultz[4], Ross L. Levine[6], Alea A. Mills[8] & Scott W. Lowe[2,9]

**Mutations disabling the *TP53* tumour suppressor gene represent the most frequent events in human cancer and typically occur through a two-hit mechanism involving a missense mutation in one allele and a 'loss of heterozygosity' deletion encompassing the other. While *TP53* missense mutations can also contribute gain-of-function activities that impact tumour progression, it remains unclear whether the deletion event, which frequently includes many genes, impacts tumorigenesis beyond *TP53* loss alone. Here we show that somatic heterozygous deletion of mouse chromosome 11B3, a 4-megabase region syntenic to human 17p13.1, produces a greater effect on lymphoma and leukaemia development than *Trp53* deletion. Mechanistically, the effect of 11B3 loss on tumorigenesis involves co-deleted genes such as *Eif5a* and *Alox15b* (also known as *Alox8*), the suppression of which cooperates with *Trp53* loss to produce more aggressive disease. Our results imply that the selective advantage produced by human chromosome 17p deletion reflects the combined impact of *TP53* loss and the reduced dosage of linked tumour suppressor genes.**

Cancer arises through the acquisition of genetic and epigenetic changes that drive tumorigenesis. While most efforts to understand the origins of these entities have focused on the identification and functional characterization of somatic single nucleotide variants (SNVs) that activate oncogenes or inactivate tumour suppressors, the vast majority of human cancers also harbour large copy number variants (CNVs) that impact gene dosage through gain or loss of whole chromosomes or chromosome segments[1]. For example, recurrent segmental deletions—such as those affecting chromosome 17p—are extremely common in human cancers yet are widely considered to arise because of selection for a single 'driver' in the deleted region, with the adjacent gene losses reflecting 'passenger' events that have no effect on tumour phenotypes. However, evidence is emerging from short hairpin RNA (shRNA) screens and bioinformatic approaches that these lesions may target more than one relevant activity[2–5]. If true, then the impact of segmental deletions on tumour development would be fundamentally distinct from SNVs, yet, given their frequent occurrence, disproportionately understudied.

## Chromosome 17p configurations in human cancer

We decided to study 17p deletion as a prototype of a recurrent cancer-associated deletion owing to its high frequency and invariable inclusion of *TP53*, an extensively studied tumour suppressor gene that is considered to be the major driver of 17p loss[6,7]. To characterize better the nature and scope of these deletions, we analysed genomic data from 4,994 tumours across 18 cancer types (http://www.cbioportal.org; accessed 29 October 2014). This survey confirmed previous indications that mutation and/or deletion of the *TP53* tumour suppressor gene occurs in approximately ~50% of all human cancers[8]. We then classified 17p-altered tumours as to whether they displayed *TP53* mutation (encompassing missense, nonsense and frameshift mutations), deletion (as defined by a reduction in *TP53* copy number), or both (Fig. 1a). The most common *TP53* configuration involves a missense mutation together with a segmental 17p deletion, although a substantial fraction of p53-altered tumours harboured chromosome 17p deletion together with an apparently wild-type *TP53* allele. Similar results were observed in blood cancers, in which *TP53* mutations and/or 17p loss are less common but are linked to a particularly dismal prognosis[9,10]. Hence, the frequency of 17p deletion may even exceed point mutations within the *TP53* gene.

Chromosome 17p deletions frequently encompass all or most of the chromosome arm (Fig. 1b). In addition to *TP53*, this region encodes over 300 protein-coding genes that include other established or putative tumour suppressors[11–13]. Unexpectedly, *TP53* was not identified as a candidate tumour suppressor in non-Hodgkin lymphoma or acute myeloid leukaemia (AML) using GISTIC, an algorithm designed to pinpoint candidate drivers from CNV data by identifying 'epicentres' of gain or loss[14] (Fig. 1b and Extended Data Fig. 1b). Moreover, AMLs harbouring a *TP53* mutation together with a 17p deletion displayed a worse prognosis than those harbouring *TP53* mutations alone (Fig. 1c and Extended Data Fig. 1c). Collectively, these data raise the possibility that additional genes drive selection for 17p loss during tumorigenesis.

## Modelling 17p13 deletions in the mouse

We decided to test this hypothesis in mice. To produce a strain capable of modelling somatic 17p deletions, we used the MICER chromosome engineering strategy developed previously[15] to introduce *loxP* sites and a split *HPRT* gene into chromosomal regions flanking the *Alox12* and *Sco1* genes on mouse chromosome 11B3, which is a 4 Mb region on mouse chromosome 11 that is syntenic to human 17p13.1 (Fig. 2a, b and Extended Data Fig. 2a, b). Candidate clones were treated with adenovirus (Adeno)-*cre* to recombine the 11B3 region, selected in

[1]Department of Hematology and Department of Liver Surgery, State Key Laboratory of Biotherapy and Cancer Center, West China Hospital, Sichuan University and National Collaborative Innovation Center, Chengdu 610041, China. [2]Department of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [3]Institute for Cancer Genetics, Columbia University Medical Center, New York, New York 10032, USA. [4]Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [5]Department of Pediatrics, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [6]Human Oncology & Pathogenesis Program and Leukemia Service, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA. [7]Department of Hematology & Research Laboratory of Hematology, West China Hospital, Sichuan University, Chengdu 610041, China. [8]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [9]Howard Hughes Medical Institute, New York, New York 10065, USA.
*These authors contributed equally to this work.

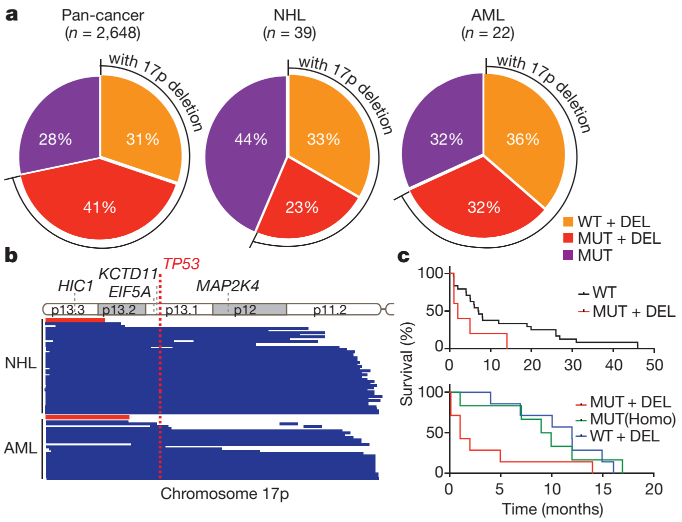


**Figure 1 | The frequency, scope, and prognostic value of chromosome 17p alterations in human cancers. a**, The nature of *TP53*-containing chromosome 17p alterations in pan-human cancer (left), non-Hodgkin lymphomas (NHL; middle) and AML (right). DEL, deletion; MUT, mutations; WT, wild-type. **b**, The extent of chromosome 17p deletions in NHL (44 cases; top) and AML (25 cases; bottom) data sets, irrespective of *TP53* deletion, as determined by single nucleotide polymorphism (SNP) array analysis. Each line represents one patient. Red bar indicates the significant copy number loss ($q < 0.25$) analysed by GISTIC algorithm. **c**, Overall survival of complex-karyotype AML patients containing both 17p deletion and *TP53* mutation as compared to those without any *TP53* alteration (top; $P = 0.076$), those with homozygous (Homo) *TP53* point mutations (bottom; $P = 0.059$), or those with 17p deletion without any detectable *TP53* mutation (bottom; $P = 0.03$). All are log-rank test.

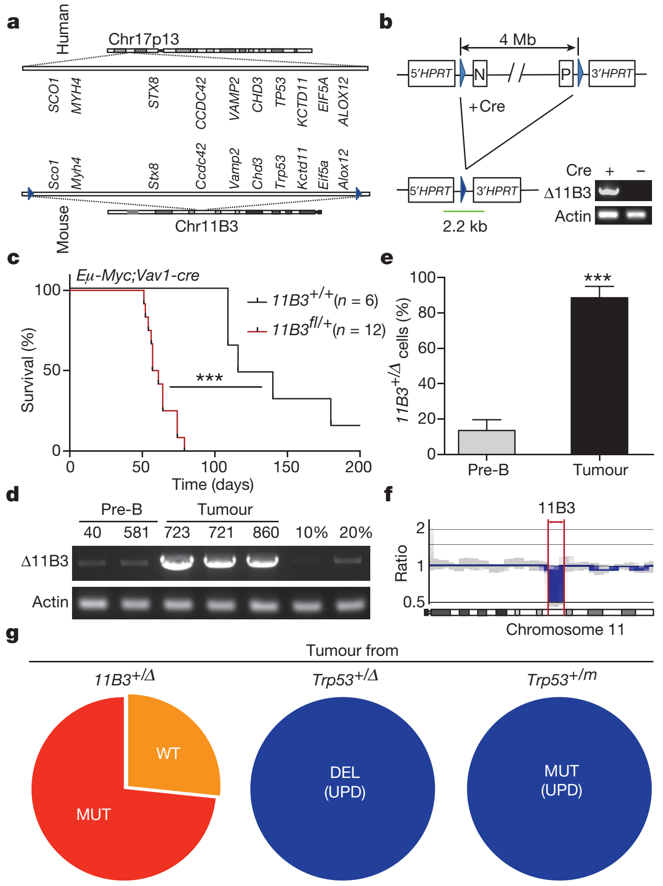


**Figure 2 | A mouse model of human 17p13.1 deletion accelerates lymphoma development. a**, The synteny of human chromosome (Chr) 17p13.1 and mouse chromosome 11B3 (from *Sco1* to *Alox12b*, ~4 Mb), with several representative genes listed. Blue arrowheads denote *loxP* sites. **b**, Conditional, 11B3-deletion strategy with PCR analysis (corresponding to the green bar) showing the desired deletion. **c**, Tumour-free survival of mice with the indicated genotype (log-rank test). **d**, **e**, The extent of 11B3 deletion in non-tumorigenic pre-B cells (*Vav1-cre;11B3*$^{fl/+}$) and in lymphomas arising in *Eμ-Myc;Vav1-cre;11B3*$^{fl/+}$ mice was determined by semi-quantitative PCR using mixed genomic DNA from *11B3*$^{+/\Delta}$ ES cells to *11B3*$^{fl/+}$ cells at different ratios (10% or 20%) (**d**), and qPCR (**e**; two-tailed *t*-test). Sample names correspond to the mouse identifier. Error bars represent standard deviation (s.d.). **f**, Copy number profile of mouse chromosome 11 as determined by low-pass whole genome sequencing of 11B3-deleted lymphoma cells obtained from **c**. Red arrows highlight the 11B3 region. **g**, Resulting *Trp53* status upon LOH from tumours arising from heterozygous chromosome 11B3 deletion (left) or those originating from heterozygous *Trp53* deletion (middle) or point mutation (right). UPD, uniparental disomy. ***$P < 0.001$.

hypoxanthine–aminopterin–thymidine (HAT) medium, and subjected to a polymerase chain reaction (PCR) analysis to discriminate between *cis* and *trans* recombination events (Extended Data Fig. 2c). This readily identified embryonic stem (ES) cell clones harbouring *loxP* sites flanking 11B3 in *cis* and revealed that 11B3 deletion is not deleterious to ES cells (data not shown). We refer to the appropriately targeted allele above as *11B3*$^{flox}$ (*11B3*$^{fl}$).

*11B3*$^{fl}$ mice were produced by blastocyst injection and crossed to a variety of constitutive and haematopoietic-directed Cre mouse lines. We never observed viable progeny harbouring a recombined 11B3 deletion arising from *11B3*$^{fl}$ crosses with *Ella-cre* or constitutive *CAG-cre* mice, implying that heterozygous deletion of the 11B3 interval confers embryonic lethality. However, double-mutant progeny harbouring the *cre* transgene and an unrecombined *11B3*$^{fl}$ allele were produced (Extended Data Table 1), perhaps surviving owing to inefficient recombination between the two distant (~4 Mb) *loxP* sites.

As in ES cells, somatic 11B3 deletion in the haematopoietic compartment was not an obligate cell lethal event; thus, recombined cells were detected in the peripheral blood of *11B3*$^{fl/+}$ mice crossed to strains harbouring *Cd19-cre*, *Mx1-cre* or *Vav1-cre* alleles (Extended Data Fig. 3a). Nevertheless, the apparent recombination frequency was substantially lower than observed in progeny of conditional *Trp53*-knockout mice crossed to the same strains (20% versus 100%; Extended Data Fig. 3b, c). While such differences can complicate direct comparisons of tumour onset between *11B3*$^{fl}$ and *Trp53*$^{fl}$ animals, we reasoned that changes in the fraction of 11B3-deleted cells over time should indicate whether 11B3 loss provides a competitive advantage during tumorigenesis.

We first tested whether somatic deletion of 11B3 could drive cancer development in the *Eμ-Myc* transgenic model of non-Hodgkin lymphoma, as *Trp53* deletion potently accelerates lymphomagenesis in this model[16,17] and because 17p alterations are associated with adverse prognosis in the corresponding human disease[18] (Extended Data Fig. 1c). Accordingly, *Eμ-Myc* mice were crossed to the *11B3*$^{fl}$ allele and a *Vav1-cre* transgenic strain, and the resulting progeny were monitored for tumour onset (Fig. 2c). Mice harbouring the *11B3*$^{fl}$ allele developed lymphomas much more rapidly than controls (59 versus 130 days, $P < 0.001$), clearly indicating that loss of the 11B3 region can promote tumorigenesis.

Paradoxically, *Eμ-Myc;Trp53*$^{fl/+}$*;Vav1-cre* mice developed lymphomas even more rapidly than *Eμ-Myc;11B3*$^{fl/+}$*;Vav1-cre* animals (Extended Data Fig. 3d). While this observation is consistent with a potentially negative impact of 11B3 deletion on cellular fitness, it might also reflect the nearly fivefold reduced recombination efficiency of the *11B3*$^{fl}$ versus *Trp53*$^{fl}$ alleles. Consistent with the idea of reduced recombination efficiency, semi-quantitative and quantitative (q)PCR analysis indicated a massive enrichment in the fraction of 11B3-deleted cells in lymphomas compared with the premalignant setting (~100% versus 15%; Fig. 2d, e), which was confirmed by genome sequencing (Fig. 2f). Although lymphomas arising in the *Trp53*$^{fl/+}$ and *11B3*$^{fl/+}$ animals were both highly disseminated, 11B3-deleted lymphomas typically presented

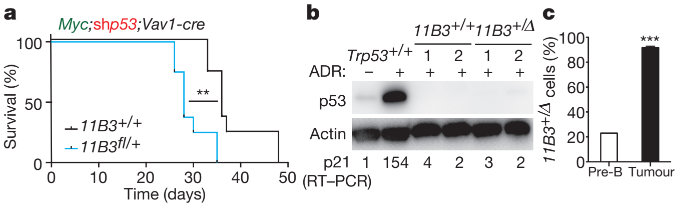


**Figure 3 | 11B3 deletion can accelerate lymphomagenesis through p53-independent mechanisms. a**, Kaplan–Meier lymphoma-free survival curve of recipient mice receiving *Vav1-cre;11B3*$^{\mathrm{fl}/+}$ or *Vav1-cre*; *11B3*$^{+/+}$, transduced simultaneously with a mouse *Myc* cDNA and an shRNA against *Trp53* (sh*p53*). $n=8$ per genotype. $^{**}P<0.01$ (log-rank test). **b**, The expression levels of *Trp53* and *p21* after ADR treatment of resulting lymphoma cells as detected by immunoblotting and RT–qPCR, respectively. *p21* levels were normalized to untreated *Trp53*$^{+/+}$ lymphomas. Shown are representative results ($n=8$ per cohort). **c**, qPCR analysis to determine the percentage of chromosome 11B3 deletion in pre-B cells (transduced cells before transplantation; $n=3$) and resulting lymphomas ($n=5$) from the experiment in **a**. $^{***}P<0.001$ (two-tailed *t*-test).

as a more mature B-cell type and showed enhanced resistance to certain chemotherapeutic agents (Extended Data Fig. 4a–c). Thus, 11B3 deletion confers a selective advantage during lymphomagenesis and produces phenotypes distinct from a *Trp53*-null setting.

p53 is induced after DNA damage, when it activates transcription of target genes involved in cell cycle arrest, apoptosis, and other cellular processes[19,20]. Strikingly, all 11B3-deleted lymphomas showed either high or undetectable levels of p53 protein, and none were capable of inducing the p53 transcriptional target *p21* (also known as *Cdkn1a*) after treatment with the DNA-damaging chemotherapeutic drug adriamycin (ADR). Such a pattern is reminiscent of lymphoma cells that acquire inactivating *Trp53* mutations[21,22] and indeed, most 11B3-deleted lymphomas (75%) acquired missense or nonsense mutations in the remaining wild-type *Trp53* allele (Extended Data Fig. 4f). Other 11B3-deleted lymphomas displayed no p53 activity, suggesting that each acquired some other mutational or epigenetic event that disabled the intact wild-type *Trp53* allele. Regardless, the majority of 11B3-deleted lymphomas display chromosome configurations analogous to the common *TP53* mutation/deletion events observed in human cancers (Fig. 1a and Fig. 2g).

We also tested the mechanisms of loss of heterozygosity (LOH) arising in conventional *Trp53*-mutant mouse models. To this end, we generated *Eμ-Myc* strains with the following additional alleles: *Trp53*$^{+/-}$, *Vav1-cre;Trp53*$^{\mathrm{fl}/+}$, *Vav1-cre;Trp53*$^{LSL\text{-}R270H/+}$, or *Vav1-cre;Trp53*$^{LSL\text{-}R172H/+}$, and analysed the fate of the remaining *Trp53* allele upon lymphoma manifestation using a qPCR assay to detect the copy number of select regions flanking *Trp53*. In these models LOH clearly does not involve deletion of the remaining *Trp53* locus, as the genes proximal and distal to *Trp53* (for example, *Eif5a* and *Sco1/Alox12*, respectively) remained diploid in all cases (Extended Data Fig. 5a). Instead, this analysis suggests that LOH occurs through duplication of the targeted mutant *Trp53* allele, a conclusion that was confirmed by analysis of polymorphic markers in this region (Fig. 2g and Extended Data Fig. 5b, c). While such 17p configurations occur in human tumours, they are less frequent than those involving large deletions (see Fig. 1a).

## Role of 11B3 deletion in lymphomagenesis

Although the data described earlier demonstrate that 11B3 deletion can contribute to tumorigenesis, the differences in recombination frequencies between the conditional 11B3 and *Trp53* alleles made it impossible to assess whether other genes in the 11B3 deletion contribute to its effects. To mitigate this caveat without extensive strain intercrossing, we exploited the capability of shRNAs to suppress gene expression in *trans*, hypothesizing that any phenotypic differences arising between wild-type and 11B3-deleted cells expressing a potent *Trp53* shRNA would be independent of *Trp53*. As immunoblotting and qPCR with reverse transcription (RT–qPCR) analysis confirmed that our *Trp53* shRNA was equally capable of disabling *Trp53* in pre-B cells harbouring one or two *Trp53* alleles (Extended Data Fig. 6), we co-transduced pre-B cells derived from the bone marrow of *Vav1-cre* and *Vav1-cre;11B3*$^{\mathrm{fl}/+}$ mice with a *Myc* complementary DNA (linked to GFP) and the *Trp53* shRNA (linked to *mCherry*) and studied tumorigenesis upon their transplantation into irradiated recipient mice.

Mice receiving cells harbouring the conditional 11B3 deletion and the *Trp53* shRNA developed tumours more rapidly than those transplanted with cells harbouring the *Trp53* shRNA alone (Fig. 3a; median onset 28 versus 36 days, $P<0.01$). In both the *11B3*$^{\mathrm{fl}/+}$ and *11B3*$^{+/+}$ settings, the disease presented as an aggressive B-cell malignancy (data not shown) in which p53 protein and activity were undetectable (Fig. 3b). Furthermore, the subsequent lymphomas were double positive for GFP and mCherry, indicating that they retained the *Myc* cDNA and *Trp53* shRNA (data not shown). While only 15–20% of the premalignant cells underwent 11B3 recombination, almost all of the resulting lymphoma cells harboured this deletion (Fig. 3c). Thus, 11B3 deletion confers a strong selective advantage even in the absence of detectable p53 function.

To identify genetic elements within the 11B3 interval that act together with *Trp53* to suppress lymphoma development, we first took a candidate gene approach. We noted that virtually all lymphoma-associated 17p deletions encompass both *TP53* and *E1F5A* (see Fig. 1b), the latter gene also having been identified as a tumour suppressor that promotes apoptosis[11]. To examine interactions between these genes during lymphomagenesis, validated shRNAs targeting *Trp53* and/or *Eif5a*, co-expressed in tandem, were introduced into *Eμ-Myc* haematopoietic stem and progenitor cells (HSPCs) and then transplanted into syngeneic recipients. As expected, suppression of either *Trp53* or *Eif5a* accelerated lymphomagenesis relative to controls, which did not develop tumours over the time of evaluation (Fig. 4a; median onset 56 and 99.5 days, respectively). Still, co-suppression of both *Trp53* and *Eif5a* produced lymphomas even more rapidly (median onset 45 days). Apparently, premalignant cells co-suppressing *Trp53* and *Eif5a* shRNAs have an increased selective advantage compared with cells expressing either shRNA alone, a suggestion that was confirmed by an *in vivo* competition assay (Extended Data Fig. 7) and apoptosis measurements in premalignant cells (Fig. 4b)[16,23,24]. Although our shRNA technology cannot precisely mimic the gene dosage produced by a heterozygous gene deletion, these data identify *Eif5a* as a second gene in the 11B3 interval that can contribute to its activity.

To test whether additional 17p13 genes have a role, we generated a shRNA library targeting the ~100 protein-coding genes in the 11B3 region (exclusive of *Trp53* and *Eif5a*) and screened it for tumour-promoting activity in HSPC transplantation assays (Fig. 4c; 17p13 shRNA library detailed in Supplementary Table 1). Compared with controls, these libraries accelerated lymphomagenesis in recipient mice, implying that one or more shRNAs conferred a selective advantage (Fig. 4d). To identify such shRNAs, PCR products amplified from genomic DNA were subject to deep sequencing and the abundance of each shRNA compared to the initial pool. Forty shRNAs targeting 17 genes were enriched in tumours compared to the injected population (Supplementary Table 1).

Among the potential candidates, two shRNAs targeting *Alox15b* (sh*Alox15b.1252* and sh*Alox15b.2865*), showed a 245- and 50-fold enrichment, respectively, compared with their representation in the initial library (Extended Data Fig. 8a). We thus tested whether suppression of *Alox15b* alone or in combination with *Trp53* could accelerate tumorigenesis as was done for *Eif5a*. Recipients of HSPCs targeted with *Alox15b* shRNAs had significantly shorter lymphoma-free survival than sh*Ren* controls (Fig. 4e), and those receiving HSPCs harbouring *Alox15b*-*Trp53* tandem shRNA showed a further acceleration over sh*Trp53*-sh*Ren* controls (Fig. 4f).

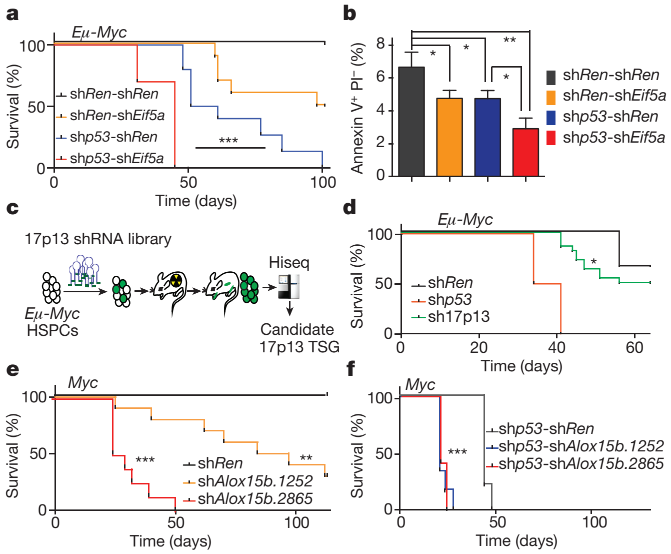


**Figure 4 | 11B3 encodes multiple genes whose attenuation cooperates with *Trp53* loss to drive lymphoma. a**, Kaplan–Meier lymphoma-free survival of recipient mice transplanted with *Eμ-Myc* HSPCs with various GFP-linked tandem shRNAs. sh*p53* indicates sh*Trp53*. $n = 10$ per group. **b**, Annexin V staining of *Eμ-Myc* pre-B cells transduced with the indicated tandem shRNAs constructs. $n = 3$. Result represents at least two independent experiments. PI, propidium iodide. **c**, **d**, *In vivo* shRNA screen to identify potential tumour suppressors on chromosome 17p13 (**c**) and resulting tumour-free survival curve of mice receiving HPSCs transduced with shRNA pools (**d**). Hiseq, high-throughput sequencing; TSG, tumour suppressor gene. **e**, **f**, Kaplan–Meier tumour-free survival curve of recipient mice transplanted with pre-B cells co-infected with *Myc* and the indicated shRNAs (**e**; $n = 10$), or infected with Myc-linked tandem shRNAs (**f**; $n = 6$). **a**, **d**–**f**, Log-rank test. **b**, Unpaired two-tailed *t*-test, error bars represent s.d. $*P < 0.05$; $**P < 0.01$; $***P < 0.001$.

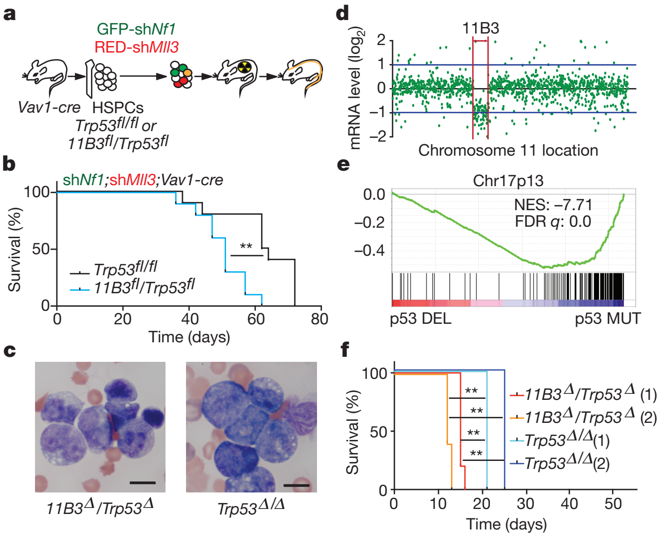


**Figure 5 | Chromosome 11B3 deletion acts beyond *Trp53* loss to drive myeloid leukaemia. a**, AML generation by HSPC isolation (*Vav1-cre;Trp53*$^{fl/fl}$ or *Vav1-cre;11B3*$^{fl}$*/Trp53*$^{fl}$), co-transduction with a GFP-linked *Nf1* shRNA and an mCherry-linked *Mll3* shRNA, and transplantation into sublethally irradiated recipients ($n = 10$ per group). **b**, Post-transplant, leukaemia-free survival $**P < 0.01$ (log-rank test). **c**, Blood smear of moribund mice (**b**) that is representative of all animals analysed in each genotype ($n = 4$). **d**, RNA-seq comparison of chromosome 11 gene-expression levels between the *11B3*$^{Δ}$*/Trp53*$^{Δ}$ and *Trp53*$^{Δ/Δ}$ leukaemia generated in **b** ($n = 4$). **e**, Gene set enrichment analysis of genes on chromosome 17p13 in chromosome 17p-deleted human AML compared to those with *TP53* mutations but without 17p deletions. FDR, false discovery rate; NES, normalized enrichment score. **f**, Kaplan–Meier survival curve of secondary transplants from two independent primary leukaemias of each genotype in **a**, **b**. $n = 5$ in each cohort. $**P < 0.01$ (log-rank test).

Partial knockdown of *Alox15b* by shRNAs in NIH3T3 cells resulted in accumulation of its substrate, arachidonic acid (AA) (Extended Data Fig. 8b, c), and higher AA levels were detected in 11B3-deleted lymphoma cells compared with their *Trp53*-null counterparts (Extended Data Fig. 8d). Interestingly, AA suppresses apoptosis in certain cancer cells[25] and, in agreement, exogenous AA inhibited apoptosis of pre-B cells in a dose-dependent manner (Extended Data Fig. 8e). While a complete mechanistic understanding of its effects will require further work, these studies pinpoint *Alox15b* as an additional tumour suppressor in the 11B3 region. Intriguingly, shRNAs targeting two other Alox family genes adjacent to *Alox15b* were enriched in our screen (*Alox12b*, *Alox3e*), raising the possibility that 17p deletion coordinately attenuates the output of this gene family. Additionally, beyond protein-coding genes, the reduced dosage of non-coding RNAs not tested in our screen might contribute to the activity of 17p deletions[26].

## 11B3 deletion contributes to AML

We also asked whether 11B3 loss would enhance tumorigenesis in a genetically and pathologically accurate mouse model of complex-karyotype AML, an invariably lethal leukaemia subtype[9,27,28]. In this disease, *TP53* lesions often co-occur with *NF1* loss and deletions on 5q and/or 7q, which can be approximated in mice transplanted with *Trp53*$^{-/-}$ HSPCs co-expressing shRNAs targeting *Nf1* and *Mll3* (also known as *Kmt2c*) (encoded on human 7q36)[29]. Taking advantage of the flexibility of this model, we compared disease onset and pathology in mice transplanted with sh*Nf1*-sh*Mll3*-transduced HSPCs derived from either *Vav1-cre;Trp53*$^{fl/fl}$ or *Vav1-cre;11B3*$^{fl}$*/Trp53*$^{fl}$ mice (Fig. 5a).

These experiments revealed a clear and substantial p53-independent effect of 11B3 deletion on leukaemia development. Indeed, despite the reduced recombination frequency of the 11B3 allele (Extended Data Fig. 9a), recipients of *Vav1-cre;11B3*$^{fl}$*/Trp53*$^{fl}$ HSPCs had a significantly decreased survival compared with recipients of *Vav1-cre;Trp53*$^{fl/fl}$ HSPCs, and succumbed to an aggressive form of AML (Fig. 5b, c and Extended Data Fig. 9b–d; median overall survival 62 versus 81 days, $P = 0.002$). RNA-sequencing (RNA-seq) analysis confirmed that 11B3 genes were underexpressed in 11B3-deleted AML cells compared with those in the *Trp53*-null setting (Fig. 5d), and gene set enrichment analysis revealed that transcriptional profiles of human 17p13 genes are downregulated in human AMLs harbouring 17p13 deletions (Fig. 5e). Moreover, 11B3-deleted AMLs were capable of serial transplantation into secondary recipients, where they promoted disease more rapidly than their *Trp53*-null counterparts (Fig. 5f). Finally, while the *Trp53*$^{-/-}$ AMLs are sensitive to the BET-protein inhibitor JQ1 (ref. 29), the corresponding *11B3*$^{-}$*/Trp53*$^{-}$ AML showed reduced sensitivity (Extended Data Fig. 9e, f). Hence, 11B3 deletions contribute p53-independent phenotypes to AML.

## Discussion

This study provides compelling evidence that 17p deletion confers phenotypes beyond those achieved through *TP53* loss alone. When considered in light of the established gain-of-function properties of certain missense *Trp53* mutations, our results imply that the most frequent somatic event in cancer contributes multiple activities that act independently of *Trp53* inactivation to drive tumorigenesis. It therefore seems likely that tumours harbouring *TP53* lesions—typically considered a uniform entity—produce distinct phenotypes owing to the nature of the *TP53* mutation and the extent of 17p deletion. As such, the models described here will be useful for dissecting precisely how different 17p configurations affect disease onset and ultimately impact therapy response. More broadly, our study provides direct *in vivo* evidence that segmental deletion events can arise owing to the selective advantage of disrupting multiple genes and provides a blueprint to dissect these common but understudied cancer-promoting lesions.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 29 April 2015; accepted 20 January 2016.**
**Published online 16 March 2016.**

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
2. Zender, L. *et al.* An oncogenomics-based *in vivo* RNAi screen identifies tumor suppressors in liver cancer. *Cell* **135,** 852–864 (2008).
3. Xue, W. *et al.* A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proc. Natl Acad. Sci. USA* **109,** 8212–8217 (2012).
4. Solimini, N. L. *et al.* Recurrent hemizygous deletions in cancers may optimize proliferative potential. *Science* **337,** 104–109 (2012).
5. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155,** 948–962 (2013).
6. Miller, L. D. *et al.* An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl Acad. Sci. USA* **102,** 13550–13555 (2005).
7. Petitjean, A., Achatz, M. I., Borresen-Dale, A. L., Hainaut, P. & Olivier, M. *TP53* mutations in human cancers: functional selection and impact on cancer prognosis and outcomes. *Oncogene* **26,** 2157–2165 (2007).
8. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6,** pl1 (2013).
9. Wattel, E. *et al.* p53 mutations are associated with resistance to chemotherapy and short survival in hematologic malignancies. *Blood* **84,** 3148–3157 (1994).
10. El-Ghammaz, A. M., Abdelwahed, E., Mostafa, N. N. & Mansour, D. A. *De novo* deletion 17p13.1 as a predictor for disease progression in chronic lymphocytic leukemia. *Clin. Exp. Med.* **15,** 493–499 (2015).
11. Scuoppo, C. *et al.* A tumour suppressor network relying on the polyamine–hypusine axis. *Nature* **487,** 244–248 (2012).
12. Wales, M. M. *et al.* p53 activates expression of *HIC-1*, a new candidate tumour suppressor gene on 17p13.3. *Nature Med.* **1,** 570–577 (1995).
13. Ahn, Y. H. *et al.* Map2k4 functions as a tumor suppressor in lung adenocarcinoma and inhibits tumor cell invasion by decreasing peroxisome proliferator-activated receptor γ2 expression. *Mol. Cell. Biol.* **31,** 4270–4285 (2011).
14. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
15. Adams, D. J. *et al.* Mutagenic insertion and chromosome engineering resource (MICER). *Nature Genet.* **36,** 867–871 (2004).
16. Schmitt, C. A. *et al.* Dissecting p53 tumor suppressor functions *in vivo*. *Cancer Cell* **1,** 289–298 (2002).
17. Eischen, C. M., Weber, J. D., Roussel, M. F., Sherr, C. J. & Cleveland, J. L. Disruption of the ARF–Mdm2–p53 tumor suppressor pathway in Myc-induced lymphomagenesis. *Genes Dev.* **13,** 2658–2669 (1999).
18. Monti, S. *et al.* Integrative analysis reveals an outcome-associated and targetable pattern of p53 and cell cycle deregulation in diffuse large B cell lymphoma. *Cancer Cell* **22,** 359–372 (2012).
19. Levine, A. J. & Oren, M. The first 30 years of p53: growing ever more complex. *Nature Rev. Cancer* **9,** 749–758 (2009).
20. Vousden, K. H. & Prives, C. Blinded by the light: the growing complexity of p53. *Cell* **137,** 413–431 (2009).
21. Olive, K. P. *et al.* Mutant p53 gain of function in two mouse models of Li–Fraumeni syndrome. *Cell* **119,** 847–860 (2004).
22. Hanel, W. *et al.* Two hot spot mutant p53 mouse models display differential gain of function in tumorigenesis. *Cell Death Differ.* **20,** 898–909 (2013).
23. Hemann, M. T. *et al.* Suppression of tumorigenesis by the p53 target PUMA. *Proc. Natl Acad. Sci. USA* **101,** 9333–9338 (2004).
24. Kelly, G. L. *et al.* Targeting of MCL-1 kills MYC-driven mouse and human lymphomas even when they bear mutations in *p53*. *Genes Dev.* **28,** 58–70 (2014).
25. Tang, D. G. *et al.* Suppression of W256 carcinosarcoma cell apoptosis by arachidonic acid and other polyunsaturated fatty acids. *Int. J. Cancer* **72,** 1078–1087 (1997).
26. Balatti, V. *et al.* *TCL1* targeting *miR-3676* is codeleted with tumor protein p53 in chronic lymphocytic leukemia. *Proc. Natl Acad. Sci. USA* **112,** 2169–2174 (2015).
27. Slovak, M. L. *et al.* Karyotypic analysis predicts outcome of preremission and postremission therapy in adult acute myeloid leukemia: a Southwest Oncology Group/Eastern Cooperative Oncology Group Study. *Blood* **96,** 4075–4083 (2000).
28. Byrd, J. C. *et al.* Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). *Blood* **100,** 4325–4336 (2002).
29. Chen, C. *et al.* MLL3 is a haploinsufficient 7q tumor suppressor in acute myeloid leukemia. *Cancer Cell* **25,** 652–665 (2014).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank J. P. Morris, L. Dow, D. Tschaharganeh, E. Manchado, T. Kitzing, E. Loizou and other Lowe laboratory members for their critical discussions and technical help, C. J. Sherr for invaluable advice, L. Dai and M. Tang for liquid chromatography–mass spectrometry support, and Y. Park and S. Kim for their help in constructing the 11B3 mouse model. This work was supported by a programme project and an R01 grant from the National Cancer Institute (S.W.L.), a Center grant for Cancer Target Discovery and Development (S.W.L.), and a Memorial Sloan Kettering Cancer Center Support Grant. Y.L. was supported by an American Association for Cancer Research Millennium Fellowship in Lymphoma Research. C.C. is supported by the Thousand Young Talents Plan and the National Natural Science Foundation of China (81522003, 81570150). S.W.L. is the Geoffrey Beene Chair for Cancer Biology and a Howard Hughes Medical Institute Investigator.

**Author Contributions** Y.L. and S.W.L. designed the experiments. Y.L., C.C., S.A., Z.X. and L.C. performed the experiments. Y.L., C.C., Z.X., T.N. and S.W.L. analysed data. Y.L., C.S. and A.A.M. designed the 11B3 model, Y.L., C.S., J.G., B.B., E.R.K., T.B., B.S., T.N., Q.W., N.S. and R.L.L. contributed to the human cancer genomic analysis. Y.L., C.C., C.D.R. and S.W.L. organized data and wrote the manuscript.

**Author Information** 17p13 shRNA library specification is provided in Supplementary Information. The raw and analysed RNA sequence data have been deposited in the Gene Expression Omnibus under accession number GSE69654. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.W.L. (lowes@mskcc.org).

## METHODS

**Genomic analyses of human cancers.** The data on *TP53* mutations (including allele frequency) and CNVs in pan-tumours and AML are derived from The Cancer Genome Atlas (TCGA) data in the cBioPortal for Cancer Genomics (http://www.cbioportal.org/; accessed on 29 October 2014). Only sequenced samples with allele frequency information provided were included in our analysis. Considering potential normal tissue contamination, samples with *TP53* mutation allele frequency above 0.6 were considered as a homozygous mutation. The SNP data were visualized in IGV and statistics for AML outcome were analysed in Prism 6. Since cBioPortal only has a few non-Hodgkin lymphoma cases available, we used published data to extract *TP53* mutation and deletion information[18,30–34]. Clinical outcomes were annotated from follow-up data available within the Gene Expression Omnibus GSE34171 series.

CNV analysis was performed using published AML and DLBCL tumour copy number data in Affymetrix SNP Array 6.0 .cel format (http://cancergenome.nih.gov/)[18,35–37] according to GISTIC2.0 (ref. 14). Specifically, the following GISTIC parameters and values were used following the latest TCGA Copy Number Portal analysis version (3 November 2014 stddata__2014_10_17; http://www.broadinstitute.org/tcga/gistic/browseGisticByTissue): core GISTIC version 2.0.22; reference genome build hg19; amplification threshold 0.1; deletion threshold 0.1; high-level amplification threshold 1.0; high-level deletion threshold 1.0; broad length cut-off 0.50; peak confidence level 0.95; cap 1.5; gene-GISTIC, true; arm-level peel-off, true; significance threshold 0.25; join segment size 8; X chromosome removed, false; maximum segments per sample 2,000; minimum samples per disease 40.

**Conditional 11B3-knockout model.** To create a conditional 11B3 chromosome deletion, the MICER strategy was used[15]. Briefly, MICER clones MHPN91j22 (centromeric to *Sco1*) and MHPP248j19 (telomeric to *Alox12*) (Sanger Institute) were introduced into AB2.2 ES cells (129S5 strain, Sanger Institute) by sequential electroporation, followed by G418 (neomycin; 180 $\mu$g ml$^{-1}$) and puromycin (1 $\mu$g ml$^{-1}$) selection, respectively. Successful recombination events were confirmed by Southern blotting using the hybridized probes designated in Supplementary Table 2 as described[38]. The *cis*- and *trans*-localizations of two *loxP* sites in doubly targeted ES cells were further distinguished by PCR with df-F and df-R, or dp-F and dp-R (Supplementary Table 2), respectively, after Adeno-*cre* infection and HAT (Gibco) selection. Correct *cis*-ES clones in which two *loxP* sites were integrated into the same allele were used to generate chimaera mice by blastocyst injection. The F1 pups were genotyped with 11B3-F and 11B3-R primers (Supplementary Table 2) and those positive backcrossed to C57BL/6 mouse strains for more than 10 generations.

**Mice.** All of the mouse experiments were approved by the Institutional Animal Care and Use Committee at the Memorial Sloan Kettering Cancer Center. *E$\mu$-Myc*, *Vav1-cre*, *Ella-cre*, *Trp53$^{LSL\text{-}R270H/+}$*, *Trp53$^{LSL\text{-}R72H/+}$*, *Trp53$^{+/-}$*, *Trp53$^{fl/+}$* and *Rag1$^{-/-}$* mice were ordered from Jackson Laboratories[21,39–44] and the *Arf$^{+/-}$* mouse strain is a gift from C. Sherr[45]. *E$\mu$-Myc* mice with different *Trp53* alterations were monitored weekly with disease state being defined by palpable enlarged solid lymph nodes and/or paralysis. Tumour monitoring was done as blinded experiments. For lymphoma generated by transplantation, 1 million *E$\mu$-Myc* HPSCs from embryonic day (E)13.5 fetal liver or autoMACS-purified B220$^{+}$ B progenitor cells isolated from 6–8-week mouse bone marrow were transduced with retroviruses, followed by tail-vein injection into sublethally irradiated (6 Gy, Cs137) C57BL/6 mice (Taconic; 6–8-week old, female, 5–10 mice per cohort)[11,46]. All recipient mice were randomly divided into subgroups before transplantation and monitored as described earlier. The generation of AML proceeded as previously reported[29]. Briefly, retrovirally infected c-Kit$^{+}$ haematopoietic stem and progenitor cells were transplanted into sublethally irradiated (6 Gy, Cs137) C57BL/6 mice, followed by routine monitoring of peripheral blood cell counts and Giemsa–Wright blood smear staining. For secondary transplantation experiments, 1 million leukaemia cells were transplanted into sublethally irradiated (4.5 Gy) mice. The immunophenotypes of resulting lymphomas and leukaemias were determined by flow cytometry as previously reported using antibodies purchased from eBioscience[11,29]. Statistical analysis of all survival data was carried out using the log-rank test from Prism 6. No statistical methods were used to predetermine sample size.

**Retroviral constructs.** MSCV-Myc-IRES-GFP and MLS-based retroviral constructs harbouring a GFP or mCherry fluorescent reporter and targeting *Ren*, *Trp53*, *Eif5a*, *Nf1* or *Mll3* have all been reported before[11,29,47]. For the tandem shRNA experiments performed in Fig. 3, mirE-based shRNAs targeting two different genes were cloned into an MLS-based vector in an analogous fashion to what has been previously described[48,49]. Retrovirus packaging and infection of HSPCs was done as previously reported[11,29].

**Apoptosis assay.** B220$^{+}$ cells were isolated from the bone marrow of 6-week-old *E$\mu$-Myc* mice by autoMACS positive selection with anti-B220 microbeads (Militeny Biotech). After overnight culture, cells were infected with retroviruses carrying the indicated shRNAs. Two days after infection, 0.5 × 10$^{6}$ cells were washed with PBS followed by annexin V buffer (10 mM HEPES, 140 mM NaCl, 25 mM CaCl$_2$, pH 7.4), and incubated at room temperature with Pacific Blue annexin V (BD Biosciences) and propridium iodide (PI; 1 $\mu$g ml$^{-1}$; Sigma-Aldrich) for 15 min and analysed on a LSR II flow cytometer (BD Biosciences). For arachidonic acid treatment, pre-B cells were cultured out from bone marrow cells in pre-B cell medium (RPMI1640, 10% FBS, 1% penicillin/streptomycin, 50 $\mu$M $\beta$-mercaptoethanol, 3 ng ml$^{-1}$ IL-7). After 3 days culture, pre-B cells were treated with a series concentration of arachidonic acid (Cayman Chemical) for 20 h, followed by annexin V staining as described earlier.

**Immunoblotting.** Lymphoma cells isolated from lymph nodes of diseased animals were treated with vehicle (PBS) or 1 $\mu$g ml$^{-1}$ adriamycin for 4 h. Whole cell lysates were extracted in cell lysis buffer (Cell Signaling Technology) supplemented with protease inhibitors (Roche), followed by SDS–PAGE gel electrophoresis and blotting onto PVDF membranes (Millipore). *E$\mu$-Myc;Arf$^{-/-}$* lymphoma cell lines were used as a positive control for p53 induction. The p53 antibody used was obtained from Novocastra (NCL-p53-505) and horseradish peroxidase (HRP)-conjugated $\beta$-actin antibody from Sigma (AC-15). Alox15b expressions were examined in NIH3T3 cells, which were infected by shRNAs targeting *Ren* or *Alox15b* and then selected by G418. Anti-Alox15b antibody is from Sigma (SAB2100110), and HRP-conjugated GAPDH antibody is from ThermoFisher Scientific (MA5-15738-HRP).

**Gene expressing profiling.** RNA-seq and data analysis were performed by the Integrated Genomic and Bioinformatics core at the Memorial Sloan Kettering Cancer Center. Briefly, total RNA from *11B3$^{fl}$/Trp53$^{fl}$*;sh*Nf1*;sh*Mll3;Vav1-cre* or *Trp53$^{fl/fl}$*;sh*Nf1*;sh*Mll3;Vav1-cre* leukaemia cells (four lines per cohort), isolated from the bone marrow of moribund mice, was isolated by Trizol extraction (Life Technologies). After ribogreen quantification (Life Technologies) and quality control on an Agilent BioAnalyzer, 500 ng of total RNA (RNA integrity number > 8) underwent polyA selection and Truseq library preparation according to instructions provided by Illumina (TruSeq RNA Sample Prep Kit v.2) with 6 cycles of PCR. Samples were barcoded and run on a Hiseq 2500 in a 50 bp/50 bp paired-end run, using the TruSeq SBS Kit v.3 (Illumina). An average of 45 million paired reads were generated per sample. At the most the ribosomal reads represented 0.1% and the percentage of mRNA bases was close to 65% on average. The output from the sequencers (FASTQ files) was mapped to the mouse genome (mm9) using the rnaStar (https://code.google.com/p/rna-star/) aligner, with the two-pass mapping methods. After mapping, the expression counts of each individual gene were computed using HTSeq (http://www-huber.embl.de/users/anders/HTSeq), followed by normalization and differential expression analysis among samples using the R/Bioconductor package DESeq (http://www-huber.embl.de/users/anders/DESeq). Gene set enrichment analysis (GSEA) was performed with Broad's GSEA algorithm.

**Quantitative PCR.** A list of all primers used for PCR analysis is given in Supplementary Table 2. For detection and quantification of 11B3 recombination/deletion two methods were employed. In both cases genomic DNA (gDNA) was extracted from lymphoma or leukaemia cells using Puregene DNA purification kit (Qiagen). Initially, semi-quantitative PCR was used to detect the recombined 11B3 allele using primers df-F and df-R, generating a 2.2 kb product (Fig. 2d). The estimated frequency of recombination was determined by dropping gDNA from *11B3$^{+/-}$* into *11B3$^{fl/+}$* at various ratios. For qPCR of the 11B3 deletion (Fig. 2e), SYBR Green PCR Master Mix (Applied Biosystems) was used and cycling and analysis was carried out on a ViiA 7 (Applied Biosystems). Primers 11B3-Q-F and 11B3-Q-R were used to detect the floxed allele, and to estimate the frequency of 11B3 deletion. Allelic frequency in UPD analysis (Extended Data Fig. 5a) was determined similarly, in this case with serial dilution of wild-type gDNA into DNase-free water to construct a standard curve. Two-tailed *t*-test is used for statistics analysis by Prism 6. For *p21* gene expression examination by RT–qPCR, RNA was isolated with Trizol, cDNA was synthesized with SuperScript III First-Strand Synthesis System (Life Technologies) and qPCR was performed as described earlier with primers p21-Q-F and p21-Q-R.

***Trp53* genomic DNA sequencing.** *Trp53* exons (2–10) were amplified from genomic DNAs of 11B3-deleted lymphomas by PCR (see Supplementary Table 2 for primer sequences) and subjected to Sanger sequencing. Mutations were called only if detected in sequencing reads carried out in the forward and reverse direction.

**SNP analysis.** SNP analysis of isolated lymphoma (tumour) or tail (normal) genomic DNAs from the same tumour-bearing mouse were carried out by Charles River laboratory. Briefly, a SNP Taqman assay with competing FAM- or VIC-labelled probes was used to detect the relevant C57BL/6 and 129S SNPs (D11Mit4 and D11NDS16) as described previously[50].

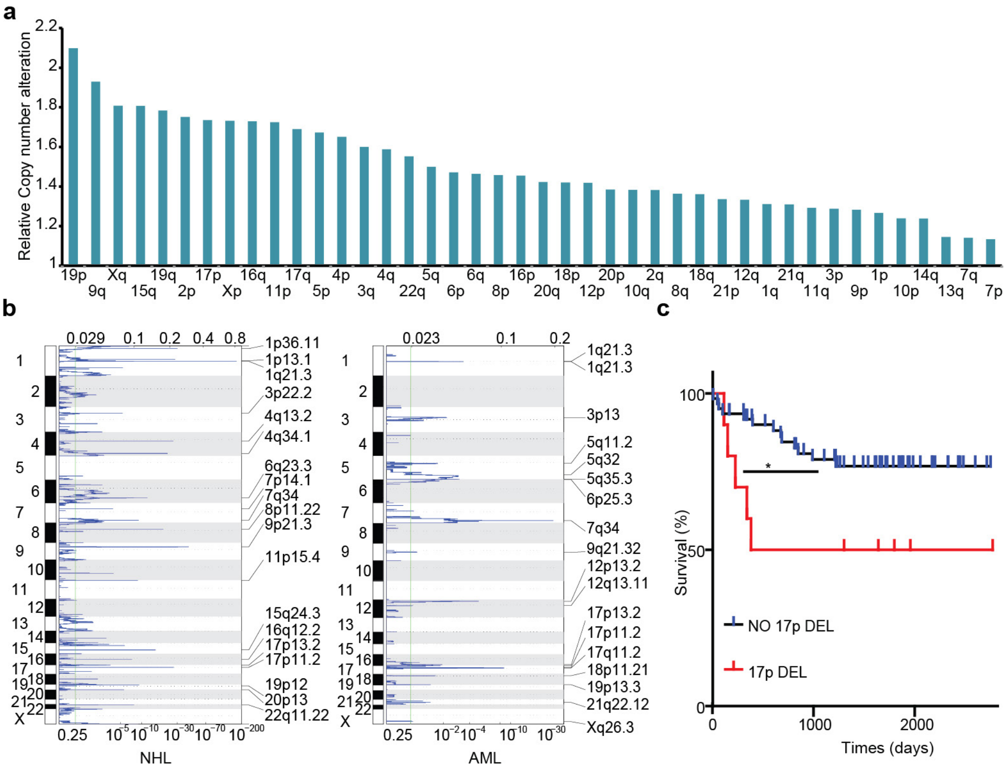
**Copy number profiling.** Genomic DNA was extracted from freshly isolated lymphoma cells from one *E$\mu$-Myc;11B3$^{fl/+}$;Vav-cre* mice. One microgram of DNA was sonicated (17 W, 75 s) on an E220 sonicator (Covaris). Samples were subsequently prepared using standard Illumina library preparation (end repair, poly A addition, and adaptor ligation). Libraries were purified using AMPure XP magnetic beads (Beckman Coulter), PCR enriched, and sequenced on an Illumina HiSeq instrument in a multiplexed format. Sequencing reads per sample were mapped using Bowtie with PCR duplicates removed. Approximately 2.5 million uniquely mappable reads were further processed for copy number determination using the 'varbin' algorithm[51,52] with 5,000 bins, allowing for a median resolution of ~600 kb. GC content normalization, segmentation and copy number estimation was calculated as described[53].

**shRNA library construction and tumour sequencing.** A custom shRNA library was designed to target mouse homologues (six shRNAs for one gene) to all human protein-coding genes on chromosome 17p13.1 from *ALOX12* to *SCO1*, except *TP53* and *EIF5A*. shRNAs were cloned into a retrovirus-based vector MLS by pool-specific PCR as previously described[11]. *Eμ-Myc* HSPCs infected with pooled shRNAs were transplanted into sublethally irradiated recipient mice. Resulting tumours were harvested, and used to extract contained shRNAs, followed by HiSeq in HiSeq 2500 (Illumina). Twenty-two oligonucleotides of shRNAs used in this study are listed in Supplementary Table 3.

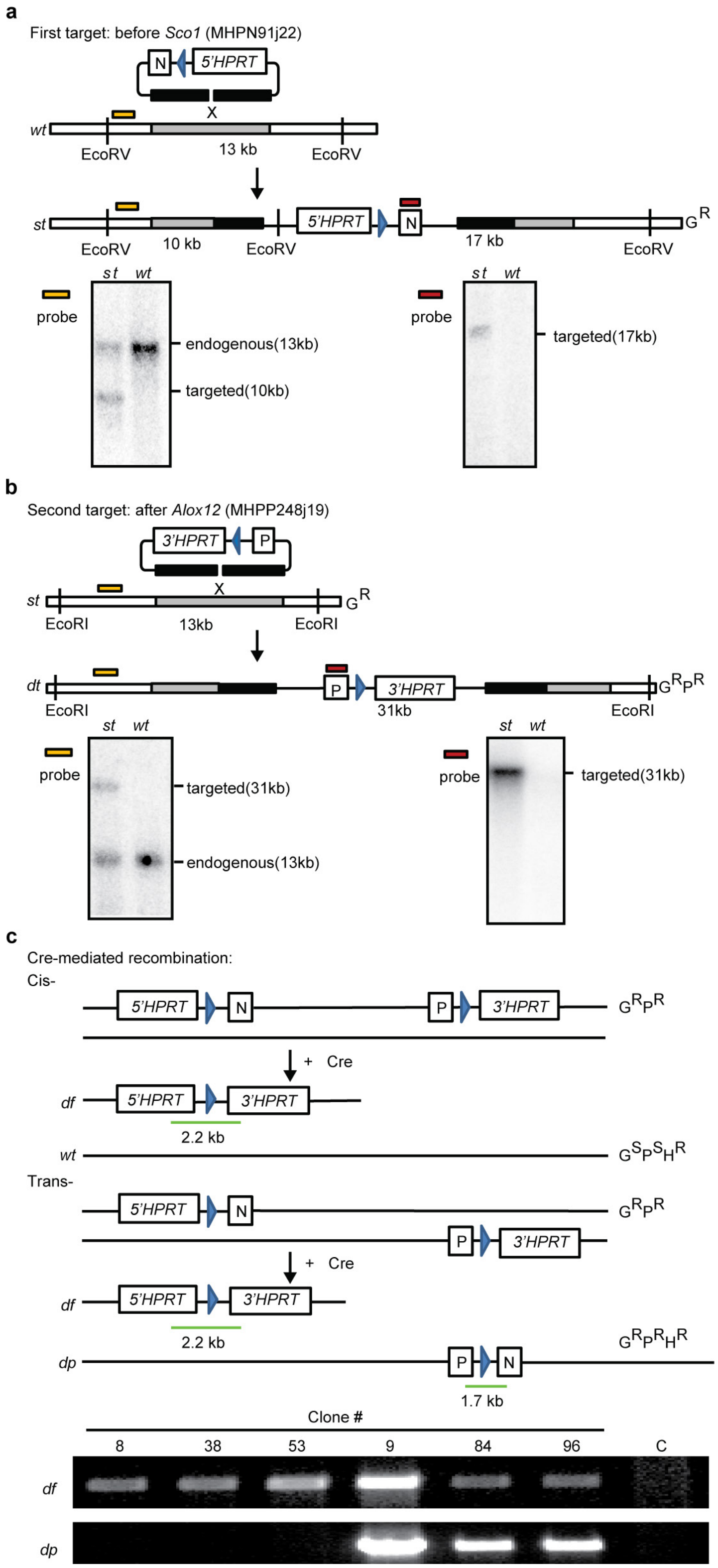
**LC-MS.** Total lipids were extracted using Folch's method[54] and analysed by LC-MS as previously described[55]. Briefly, freshly harvested cells were homogenized by chloroform/methanol (2:1, v-v). After being washed by water, the lipid-containing chloroform phase is evaporated. Dried lipids were dissolved in 100 μl 95% acetonitrile (in $H_2O$), sonicated for 3–5 min, and spiked with 10 μl of 500 ng ml$^{-1}$ deuterated internal standard solution (IS; arachidonic acid-d8; Cayman Chemical, 390010). Then, 5 μl samples were injected into Acquity ultra performance liquid chromatography (UPLC) system (Waters), equipped with Acquity UPLC BEH C18 column (100 mm × 2.1 mm I.D., 1.7 μm; Waters). Samples were washed through the column with a gradient 0.1% formic acid: acetonitrile mobile elution from 35:65 (v:v) to 5:95 for 10 min. Flow rate was 0.25 ml min$^{-1}$. Right after HPLC, samples were analysed in a Quattro Premier EX triple quadrupole mass spectrometer (Waters), which has electrospray negative mode and MasslynxV4.1 software. For each run, a standard curve was generated with different concentration of arachidonic acid lipid maps MS standard (Cayman Chemical, 10007268) mixed with IS (50 ng ml$^{-1}$ final concentration). Arachidonic acid standard $m/z$ is 303.2, and IS is 311.3.

***In vitro* drug response assays.** Three *Eμ-Myc* lymphoma cell lines generated from *Trp53*$^{fl/+}$*;Vav1-cre* or *11B3*$^{fl/+}$*;Vav1-cre* tumour-bearing mice were cultured in BCM medium (45% DMEM, 45% IMDM, 10% FBS, 2 mM glutamine, 50 μM β- mercaptoethanol, 1× penicillin/streptomycin) in 96-well plates. Cells were treated with the indicated concentrations of 4-hydroxycyclophosphamide (Toronto Research Chemicals) or vincristine (Bedford Laboratories) for 3 days. The number of living cells was determined by PI staining and cell counting on a Guava EasyCyte (EMD Millipore). Leukaemia cell lines from *Trp53*$^{\Delta/\Delta}$ or *11B3*$^{\Delta}$/*Trp53*$^{\Delta}$;sh*Nf1*;sh*Mll3* mice were treated with cytarabine (araC; Bedford Laboratories) or JQ1 (a gift from J. Bradner) in stem cell medium (BCM medium supplemented with 1 ng ml$^{-1}$ IL-3, 4 ng ml$^{-1}$ IL-6 and 10 ng ml$^{-1}$ SCF) and cell viability after 3 days was determined similarly. All cytokines are from Invitrogen.

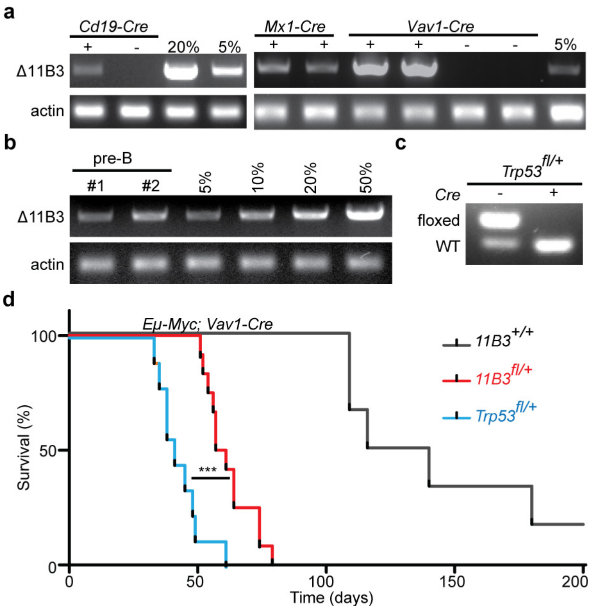
30. Salaverria, I. *et al.* Specific secondary genetic alterations in mantle cell lymphoma provide prognostic information independent of the gene expression-based proliferation signature. *J. Clin. Oncol.* **25,** 1216–1222 (2007).
31. Rubio-Moscardo, F. *et al.* Mantle-cell lymphoma genotypes identified with CGH to BAC microarrays define a leukemic subgroup of disease and predict patient outcome. *Blood* **105,** 4445–4454 (2005).
32. Chen, W. *et al.* Array comparative genomic hybridization reveals genomic copy number changes associated with outcome in diffuse large B-cell lymphomas. *Blood* **107,** 2477–2485 (2006).
33. Bea, S. *et al.* Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood* **106,** 3183–3190 (2005).
34. Mestre-Escorihuela, C. *et al.* Homozygous deletions localize novel tumor suppressor genes in B-cell lymphomas. *Blood* **109,** 271–280 (2007).
35. Rücker, F. G. *et al.* *TP53* alterations in acute myeloid leukemia with complex karyotype correlate with specific copy number alterations, monosomal karyotype, and dismal outcome. *Blood* **119,** 2114–2121 (2012).
36. Chigrinova, E. *et al.* Two main genetic pathways lead to the transformation of chronic lymphocytic leukemia to Richter syndrome. *Blood* **122,** 2673–2682 (2013).
37. The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368,** 2059–2074 (2013).
38. Brown, T. Southern blotting. Curr. Protoc. *Prot. Sci. 13,* 4G:A.4G.1–4G:A.4G.8 (2001).
39. Adams, J. M. *et al.* The c-*myc* oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature* **318,** 533–538 (1985).
40. de Boer, J. *et al.* Transgenic mice with hematopoietic and lymphoid specific expression of Cre. *Eur. J. Immunol.* **33,** 314–325 (2003).
41. Lakso, M. *et al.* Efficient *in vivo* manipulation of mouse genomic sequences at the zygote stage. *Proc. Natl Acad. Sci. USA* **93,** 5860–5865 (1996).
42. Jacks, T. *et al.* Tumor spectrum analysis in *p53*-mutant mice. *Curr. Biol.* **4,** 1–7 (1994).
43. Marino, S. & Vooijs, M., van Der Gulden, H., Jonkers, J. & Berns, A. Induction of medulloblastomas in *p53*-null mutant mice by somatic inactivation of *Rb* in the external granular layer cells of the cerebellum. *Genes Dev.* **14,** 994–1004 (2000).
44. Mombaerts, P. *et al.* RAG-1-deficient mice have no mature B and T lymphocytes. *Cell* **68,** 869–877 (1992).
45. Kamijo, T. *et al.* Tumor suppression at the mouse *INK4a* locus mediated by the alternative reading frame product p19$^{ARF}$. *Cell* **91,** 649–659 (1997).
46. Chien, Y. *et al.* Control of the senescence-associated secretory phenotype by NF-κB promotes senescence and enhances chemosensitivity. *Genes Dev.* **25,** 2125–2136 (2011).
47. Hemann, M. T. *et al.* Evasion of the p53 tumour surveillance network by tumour-derived MYC mutants. *Nature* **436,** 807–811 (2005).
48. Fellmann, C. *et al.* An optimized microRNA backbone for effective single-copy RNAi. *Cell Reports* **5,** 1704–1713 (2013).
49. Chicas, A. *et al.* Dissecting the unique role of the retinoblastoma tumor suppressor during cellular senescence. *Cancer Cell* **17,** 376–387 (2010).
50. Simpson, E. M. *et al.* Genetic variation among 129 substrains and its importance for targeted mutagenesis in mice. *Nature Genet.* **16,** 19–27 (1997).
51. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472,** 90–94 (2011).
52. Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nature Protocols* **7,** 1024–1041 (2012).
53. Baslan, T. & Hicks, J. Single cell sequencing approaches for complex biological systems. *Curr. Opin. Genet. Dev.* **26,** 59–65 (2014).
54. Folch, J., Lees, M. & Sloane Stanley, G. H. A simple method for the isolation and purification of total lipides from animal tissues. *J. Biol. Chem.* **226,** 497–509 (1957).
55. Ye, X. *et al.* Development and validation of a UPLC-MS/MS method for quantification of SKLB010, an investigational anti-inflammatory compound, and its application to pharmacokinetic studies in beagle dogs. *J. Pharm. Biomed. Anal.* **56,** 366–372 (2011).



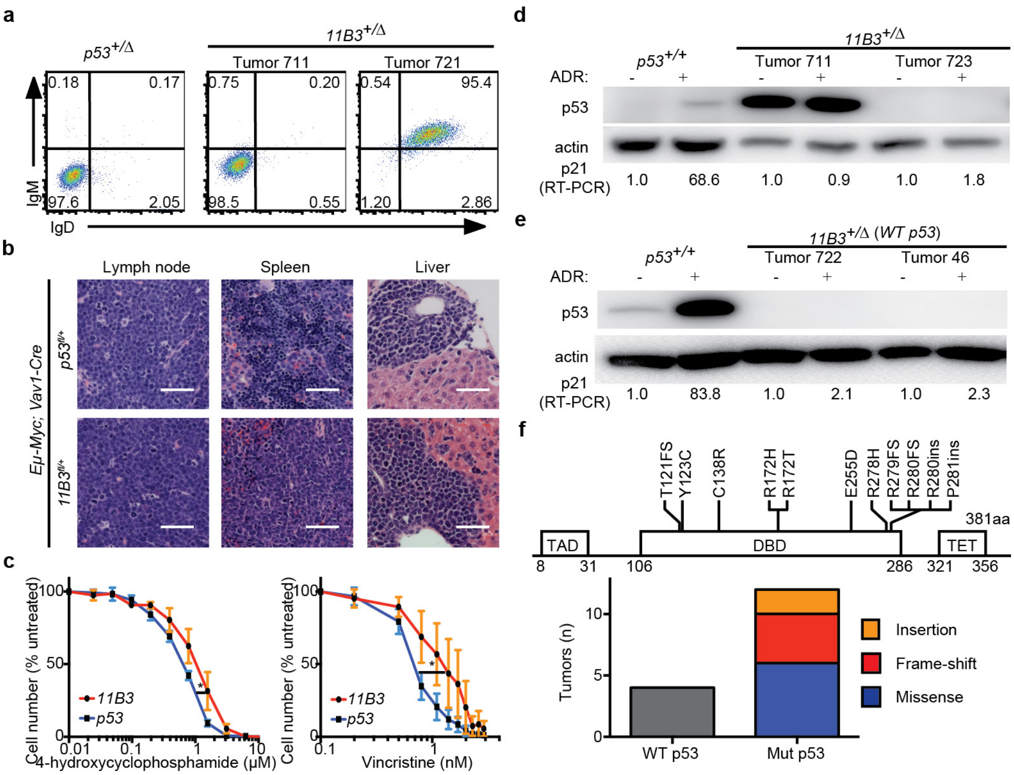
**Extended Data Figure 1 | The frequency and prognostic impact of chromosome 17p deletion with the copy number loss identified by GISTIC. a**, The ratio of chromosome-copy-number-altered cases within *TP53*-mutated cases compared to those within wild-type cases. *TP53* mutations were statistically correlated with 17p loss ($P < 0.001$) but also other copy number events ($P < 0.001$). **b**, Peaks of copy number loss identified by the GISTIC algorithm in NHL or AML. *x* axis, GISTIC *q* value; *y* axis, chromosome. $q < 0.25$ is considered as significant. **c**, Overall survival of human diffuse large B-cell lymphoma (DLBCL) patients with chromosome 17p deletion is significantly shortened compared to those with no 17p copy number variants, as annotated from the Gene Expression Omnibus GSE34171 series. *$P < 0.05$ (log-rank test).



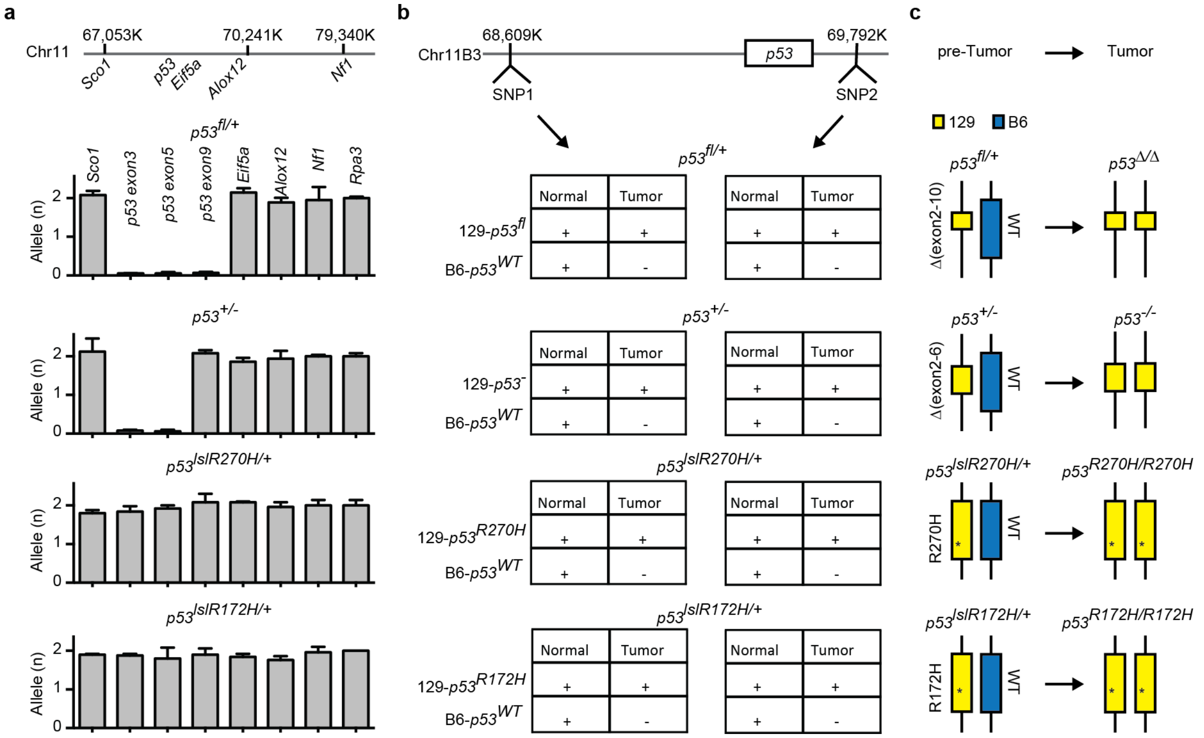
**Extended Data Figure 2 | Generation of a chromosome 11B3 conditional knockout mouse.** **a**, Top, strategy to introduce 5′ *HPRT* gene and *loxP* site telomeric to *Sco1* on chromosome 11B3 with MICER clone MHPN91j22. Bottom, Southern blot demonstrating correct targeting of the derived ES cells. *st*, single-targeted allele; *wt*, wild-type. Blue arrowheads denote *loxP* sites. **b**, Top, strategy to introduce 3′ *HPRT* gene and *loxP* site centromeric to *Alox12* on chromosome 11B3 with MICER clone MHPP248j19. Bottom, Southern blot demonstrating correct targeting of the derived ES cells. dt, double-targeted allele. **c**, Top, diagram showing the expected PCR results and drug-resistance phenotypes of doubly targeted ES cells harbouring *loxP* sites in *cis* versus in *trans*. G$^R$, G418 (neomycin) resistance; P$^R$, puromycin resistance; H$^R$, HAT resistance. *df*, deleted allele; *dp*, duplicated allele. Green bar indicates the PCR product location and length. Bottom, PCR results show different ES cell clones generated in **a** and **b**.



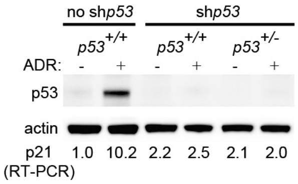
**Extended Data Figure 3 | 11B3 recombination and lymphomagenesis in *Eu-Myc* model. a**, The extent of 11B3 deletion in peripheral blood cells, as determined by semi-quantitative PCR, in *11B3*$^{fl/+}$ mice crossed to *Cd19-cre* (left), *Mx1-cre* (middle) or *Vav1-cre* (right). Genomic DNA from *11B3*$^{+/\Delta}$ ES cells was mixed with *11B3*$^{fl/+}$ cells at different ratios (5% or 20%) as a standard. For *Mx1-cre*, 6–8-week-old mice were treated with polyinosinic:polycytidylic acid (poly(I:C)) (15 mg kg$^{-1}$ every other day, 7 times) by intraperitoneal injection. **b**, Partial 11B3 deletion in *Vav1-cre;11B3*$^{fl/+}$ pre-B cells as determined by semi-quantitative PCR, indicating incomplete recombination. **c**, Complete *Trp53*$^{fl/+}$ recombination in *Vav1-cre;Trp53*$^{fl/+}$ pre-B cells as determined by PCR. **d**, Tumour-free survival of *Eμ-Myc;Vav1-cre;Trp53*$^{fl/+}$ ($n = 9$), *Eμ-Myc;Vav1-cre;11B3*$^{fl/+}$ ($n = 12$) and *Eμ-Myc;Vav1-cre* ($n = 6$) mice shows that 11B3-deleted tumours have longer tumour latency than *Trp53*-loss-only controls. ***$P < 0.001$ (log-rank test).



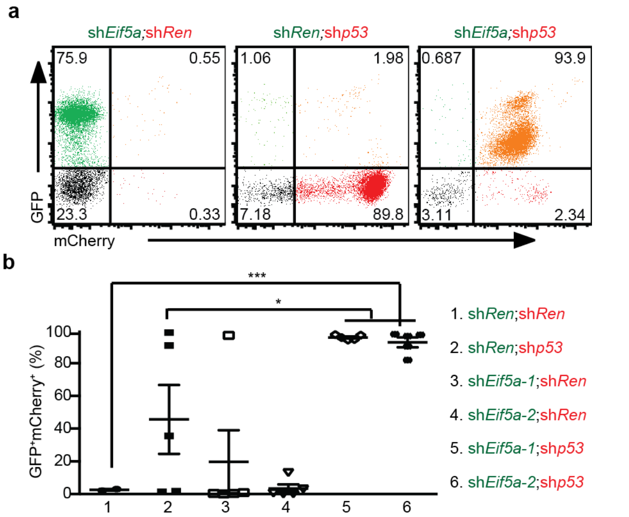
**Extended Data Figure 4 | Charaterization of 11B3-deleted lymphoma.** **a**, Immunophenotypes of B220$^+$ *Eμ-Myc* lymphomas generated from *Vav1-cre;p53*$^{fl/+}$ or *Vav1-cre;11B3*$^{fl/+}$. 11B3-deleted lymphomas were either IgM$^-$IgD$^-$ or IgM$^+$IgD$^+$ while all the *Trp53*-null lymphomas were IgM$^-$IgD$^-$. **b**, Haematoxylin and eosin (H&E) stainings of lymph node, spleen and liver of moribund, lymphoma-bearing mice originating from *Eμ-Myc;Vav1-cre;11B3*$^{fl/+}$ or *Eμ-Myc;Vav1-cre;Trp53*$^{fl/+}$ genotypes. Scale bar, 50 μm. **c**, 11B3-deleted lymphoma cells isolated from enlarged lymph nodes are more resistant to chemotherapy drugs 4-hydroxycyclophosphamide (left) and vincristine (right), by *in vitro* drug sensitivity assay. Shown are representative results of three *11B3*$^{\Delta}$/*Trp53*$^{\Delta frameshift}$ (*11B3*) or *Trp53*$^{\Delta/\Delta}$ (*p53*) lymphoma cell lines assayed in quadruplicate. *$P < 0.05$ (Student's two-tailed *t*-test). **d**, **e**, No functional p53 was detected in various 11B3-deleted tumours as determined by western blotting of p53 and RT–PCR analysis of *p21* induction after 4-h ADR treatment. *Eμ-Myc;Arf*$^{-/-}$ (*Trp53*$^{+/+}$) lymphomas were used as a positive control and *p21* levels were normalized to untreated cells. Tumours shown in **d** were identified as missense (tumour 711) or frameshift mutations (tumour 723), while those in **e** had no detectable mutation. In total eight tumours were analysed. **f**, The scope of p53 mutations detected in chromosome 11B3-deleted lymphoma cells as determined by sequencing ($n = 12$). DBD, DNA-binding domain; FS, frameshift mutation; INS, insertion mutation; MS, missense mutation; TAD, transactivation domain; TET, tetramerization domain.



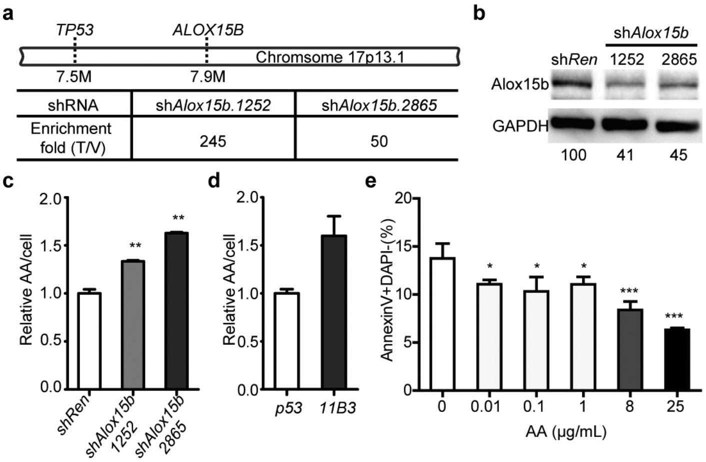
**Extended Data Figure 5 | Tumours in mice heterozygous for *Trp53* mutations lose heterozygosity by duplicating the mutant *Trp53* allele.** **a**, No chromosome 11B3 deletion was detected in various *Trp53* heterozygous mutants. Relative allele copy number of various chromosome 11B3 genes, as determined by qPCR analysis of genomic DNA from *E$\mu$-Myc* lymphomas derived from germline mice harbouring the following additional alleles: *Vav1-cre;$Trp53^{fl/+}$*(exon 2–10 flanked), *$Trp53^{+/-}$* (exon 2–6 deleted), *Vav1-cre;$Trp53^{LSL-R270H/+}$* or *Vav1-cre;$Trp53^{LSL-R172H/+}$*. *Rpa3* on chromosome 6 was used as an endogenous normalization control. **b**, SNP analysis of tumour or normal tissue (tail) genomic DNA harvested from mice in **a**, indicating that uniparental disomy occurred during *Trp53* LOH, in that C57BL/6 (B6)-derived wild-type *Trp53* allele is replaced by 129-derived *Trp53* mutant allele. Note that all *Trp53*-engineered alleles retain 129-derived SNPs; the germline wild-type *Trp53* allele is C57BL/6-derived. **c**, Cartoon summary of the results from **a** and **b**.



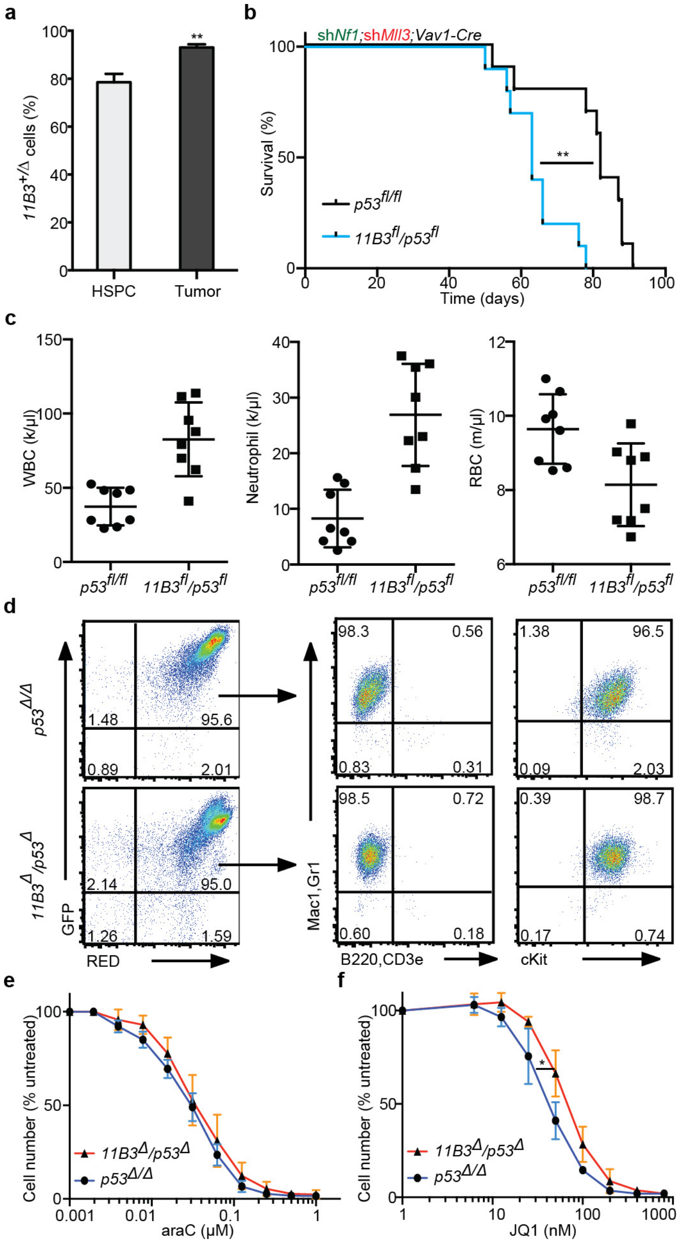
**Extended Data Figure 6 | A *Trp53* shRNA induces equivalent knockdown in cells with one or two alleles of the *Trp53* gene.** Pre-B cells were isolated from *Trp53*$^{+/+}$ or *Trp53*$^{+/-}$ bone marrow, and then transduced with GFP-linked *Trp53* shRNA (sh*p53*). GFP$^{+}$ cells were sorted out by fluorescence-activated cell sorting, and treated with control wild-type pre-B cells in the present of vehicle or 1 μg ml$^{-1}$ ADR for 4 h. p53 and *p21* levels were detected by western blotting and RT–qPCR, respectively. Shown is the representative result of three independent experiments.



**Extended Data Figure 7 | In the *Eμ-Myc* model, *Trp53* and *Eif5a* cooperate in tumorigenesis.** Two-colour assay for the cooperation of *Trp53* and *Eif5a* deficiencies on lymphoma genesis. *Eμ-Myc* HSPCs retrovirally co-transduced with *GFP*- (sh*Eif5a* or sh*Ren*) and *mCherry*-linked shRNAs (sh*Ren*, sh*p53*) were transplanted into sublethally-irradiated syngeneic recipients ($n = 5$ per group). **a**, **b**, The resulting tumours were analysed by flow cytometry (**a**) and the percentage of GFP$^+$mCherry$^+$ lymphoma cells in each configuration was quantified (**b**). Error bars represent s.d. *$P < 0.05$, ***$P < 0.001$ (two tailed $t$-test).



**Extended Data Figure 8 | *Alox15b* deficiency promotes tumorigenesis and increases AA levels. a**, Enrichment fold of sh*Alox15b.1252* and sh*Alox15b.2865* in resulting tumours (Fig. 3i, j) compared to those in initiating shRNA libraries. **b**, Knockdown efficiency of sh*Alox15b.1252* and sh*Alox15b.2865* compared to control sh*Ren* in NIH3T3 cells, as detected by western blotting and quantitated by ImageJ. **c**, Relative levels of AA per cell are increased with *Alox15b* shRNAs as measured by liquid chromatography–mass spectrometry (LC-MS). NIH3T3 cells were transduced with sh*Ren* or sh*Alox15b*. $n = 3$. $^{**}P < 0.01$ (unpaired two tailed *t*-test). **d**, Relative levels of AA per cell in $11B3^{\Delta}/Trp53^{\Delta frameshift}$ (*11B3*) lymphoma cells are higher than control cells with $Trp53^{\Delta/\Delta}$ (*p53*) as measured by LC-MS. $n = 2$. $P = 0.056$ (unpaired two tailed *t*-test). **e**, *In vitro* AA treatments reduce apoptosis, as measured by annexin V staining of pre-B cells after 20 h treatment of indicated concentration of AA. $n = 4$. $^{*}P < 0.05$; $^{***}P < 0.001$ (unpaired two tailed *t*-test).



**Extended Data Figure 9 | 11B3 deletion accelerates leukaemogenesis beyond *Trp53* loss alone and decreases sensitivity to the BET-protein inhibitor JQ-1. a**, The percentage of 11B3 deletion as determined by qPCR in premalignant c-Kit$^+$ HSPCs ($n = 2$) and resulting leukaemia cells (tumour; $n = 4$). $^{**}P < 0.01$ (unpaired two-tailed *t*-test). **b**, Overall survival of recipient mice transplanted with HSPCs from *Vav1-cre;11B3*$^{fl}$/*Trp53*$^{fl}$ or *Vav1-cre;Tr53*$^{fl/fl}$ co-transduced with both *Nf1* and *Mll3* shRNAs. $^{**}P < 0.01$ (log-rank test). **c**, Complete blood cell counts of recipient mice indicate that there are more total white blood cells (WBCs) and neutrophils, and fewer red blood cells in *Vav1-cre;11B3*$^{fl}$/*Trp53*$^{fl}$ mice compared with the *Vav1-cre;Trp53*$^{fl/fl}$ control group at 8 weeks post-transplantation. (Note that two mice from each group died before analysis and were not included.) **d**, Flow cytometry analysis of GFP$^+$mCherry$^+$ leukaemic cells in the bone marrow of moribund mice in **a** shows that leukaemia cells are myeloid cells in origin and contain both sh*Nf1* and sh*Mll3*. **e, f**, *In vitro* drug sensitivity of leukaemia cells to araC (**e**) and the BET-bromodomain inhibitor JQ-1 (**f**). Shown are representative results of three *11B3*$^{Δ}$/*Trp53*$^{Δ}$ and two *Trp53*$^{Δ/Δ}$ leukaemia cell lines assayed in quadruplicate. $^{*}P < 0.05$ (Student's two-tailed *t*-test).

**Extended Data Table 1 | Number and genotype of progeny resulting from crosses of *11B3*$^{fl/+}$*;Ella-cre* female mice with *Ella-cre* male mice**

| Progeny genotype | *11B3*$^{+/\Delta}$ | *11B3*$^{+/+}$ |
|---|---|---|
| Predicted | 47 | 47 |
| Observed | 0* | 69 |

No 11B3-deleted pup was produced from *11B3*$^{fl/+}$ female breeders with germline *Ella-cre*, while 25 out of 94 pups were genotyped as *11B3*$^{fl/+}$ without recombination.

*$P = 2.8 \times 10^{-7}$ (Chi-squared test).

doi:10.1038/nature17147

# Acceleration of petaelectronvolt protons in the Galactic Centre

HESS Collaboration*

**Galactic cosmic rays reach energies of at least a few petaelectronvolts[1] (of the order of $10^{15}$ electronvolts). This implies that our Galaxy contains petaelectronvolt accelerators ('PeVatrons'), but all proposed models of Galactic cosmic-ray accelerators encounter difficulties at exactly these energies[2]. Dozens of Galactic accelerators capable of accelerating particles to energies of tens of teraelectronvolts (of the order of $10^{13}$ electronvolts) were inferred from recent γ-ray observations[3]. However, none of the currently known accelerators—not even the handful of shell-type supernova remnants commonly believed to supply most Galactic cosmic rays—has shown the characteristic tracers of petaelectronvolt particles, namely, power-law spectra of γ-rays extending without a cut-off or a spectral break to tens of teraelectronvolts[4]. Here we report deep γ-ray observations with arcminute angular resolution of the region surrounding the Galactic Centre, which show the expected tracer of the presence of petaelectronvolt protons within the central 10 parsecs of the Galaxy. We propose that the supermassive black hole Sagittarius A* is linked to this PeVatron. Sagittarius A* went through active phases in the past, as demonstrated by X-ray outbursts[5] and an outflow from the Galactic Centre[6]. Although its current rate of particle acceleration is not sufficient to provide a substantial contribution to Galactic cosmic rays, Sagittarius A* could have plausibly been more active over the last $10^6$–$10^7$ years, and therefore should be considered as a viable alternative to supernova remnants as a source of petaelectronvolt Galactic cosmic rays.**

The large photon statistics accumulated over the last 10 years of observations with the High Energy Stereoscopic System (HESS), together with improvements in the methods of data analysis, allow for a deep study of the properties of the diffuse very-high-energy (VHE; more than 100 GeV) emission of the central molecular zone. This region surrounding the Galactic Centre contains predominantly molecular gas and extends (in projection) out to radius $r \approx 250$ pc at positive Galactic longitudes and $r \approx 150$ pc at negative longitudes. The map of the central molecular zone as seen in VHE γ-rays (Fig. 1) shows a strong (although not linear; see below) correlation between the brightness distribution of VHE γ-rays and the locations of massive gas-rich complexes. This points towards a hadronic origin of the diffuse emission[7], where the γ-rays result from the interactions of relativistic protons with the ambient gas. The other important channel of production of VHE γ-rays is the inverse Compton (IC) scattering of electrons. However, the severe radiative losses suffered by multi-TeV electrons in the Galactic Centre region prevent them from propagating over scales comparable to the size of the central molecular zone, thus disfavouring a leptonic origin of the γ-rays (see discussion in Methods and Extended Data Figs 1 and 2).

The location and the particle injection rate history of the cosmic-ray accelerator(s) responsible for the relativistic protons determine the spatial distribution of these cosmic rays which, together with the gas distribution, shape the morphology of the central molecular zone seen in VHE γ-rays. Figure 2 shows the radial profile of the $E \geq 10$ TeV cosmic-ray energy density $w_{\mathrm{CR}}$ up to $r \approx 200$ pc (for a Galactic Centre distance of 8.5 kpc), determined from the γ-ray luminosity and the amount of target gas (see Extended Data Tables 1 and 2). This high energy density in the central molecular zone is found to be an order of magnitude larger than that of the 'sea' of cosmic rays that universally fills the Galaxy, while the energy density of low energy (GeV) cosmic rays in this region has a level comparable to it[8]. This requires the presence of one or more accelerators of multi-TeV particles operating in the central molecular zone.

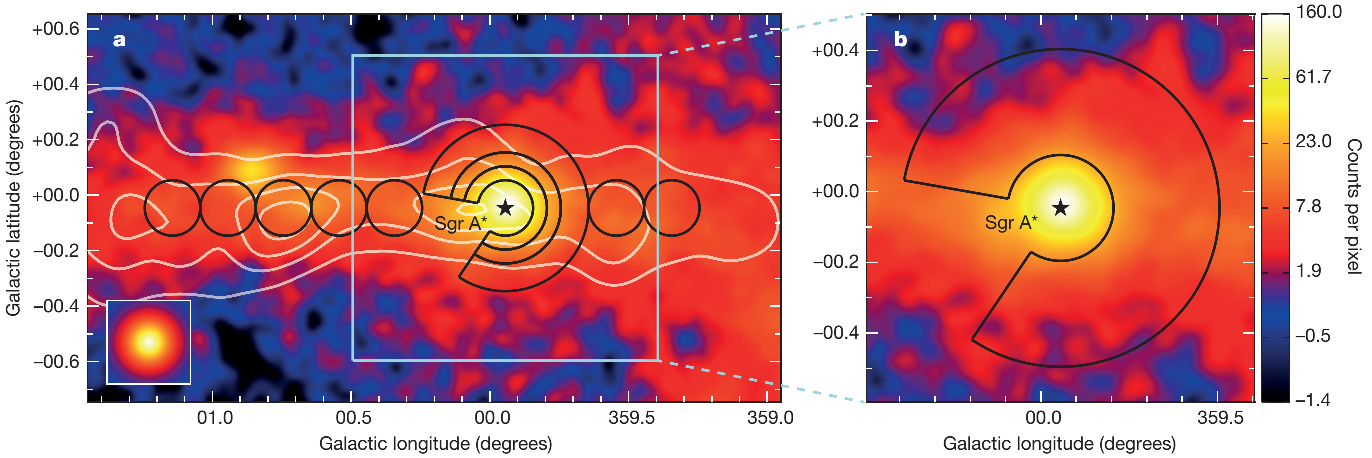


**Figure 1 | VHE γ-ray image of the Galactic Centre region.** The colour scale indicates counts per 0.02° × 0.02° pixel. **a**, The black lines outline the regions used to calculate the cosmic-ray energy density throughout the central molecular zone. A section of 66° is excluded from the annuli (see Methods). White contour lines indicate the density distribution of molecular gas, as traced by its CS line emission[30]. Black star, location of Sgr A*. Inset (bottom left), simulation of a point-like source. The part of the image shown boxed is magnified in **b**. **b**, Zoomed view of the inner ~70 pc and the contour of the region used to extract the spectrum of the diffuse emission.

*Lists of participants and their affiliations appear at the end of the paper.

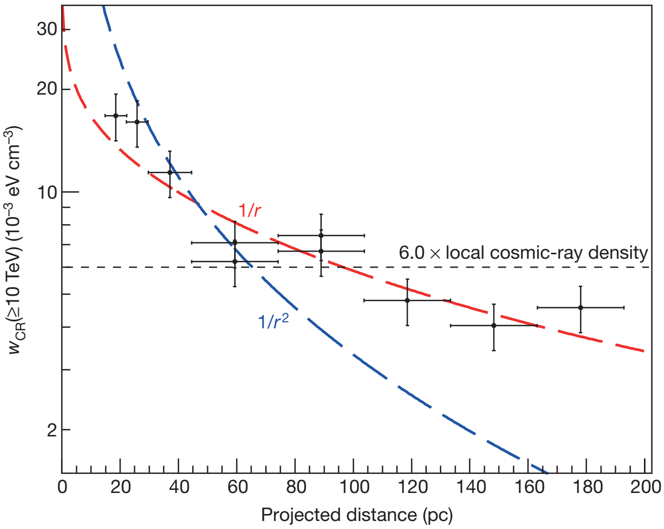


**Figure 2 | Spatial distribution of the cosmic-ray density versus projected distance from Sgr A*.** The vertical and horizontal error bars show the 1$\sigma$ statistical plus systematic errors and the bin size, respectively. Fits to the data of a $1/r$ (red line, $\chi^2$/d.o.f. = 11.8/9), a $1/r^2$ (blue line, $\chi^2$/d.o.f. = 73.2/9) and a homogeneous (black line, $\chi^2$/d.o.f. = 61.2/9) cosmic-ray density radial profile integrated along the line of sight are shown. The best fit of a $1/r^{\alpha}$ profile to the data is found for $\alpha = 1.10 \pm 0.12$ (1$\sigma$). The $1/r$ radial profile is clearly preferred for the HESS data.

If the accelerator injects particles (here we consider protons throughout) at a continuous rate, $\dot{Q}_p(E)$, the radial distribution of cosmic rays in the central molecular zone, in the case of diffusive propagation, is described[9] as $w_{\rm CR}(E,r,t) = \frac{\dot{Q}_p(E)}{4\pi D(E) r}{\rm erfc}(r/r_{\rm diff})$, where $D(E)$ and $r_{\rm diff}$ are the diffusion coefficient and radius, respectively. For timescales $t$ smaller than the proton–proton interaction time ($t_{pp} \approx 5\times 10^4 (n/10^3)^{-1}$ yr, where $n$ is the density of the hydrogen gas in cm$^{-3}$), the diffusion radius is $r_{\rm diff} \approx \sqrt{4D(E)t}$. Thus, at distances $r < r_{\rm diff}$, the proton flux should decrease as $\sim 1/r$ provided that the diffusion coefficient does not vary much throughout the central molecular zone. The measurements clearly support the $w_{\rm CR}(r) \propto 1/r$ dependence over the entire central molecular zone region (Fig. 2) and disfavour both $w_{\rm CR}(r) \propto 1/r^2$ and $w_{\rm CR}(r) \propto$ constant profiles (the former is expected if cosmic rays are advected in a wind, and the latter in the case of a single burst-like event of cosmic-ray injection). The $1/r$ profile of the cosmic-ray density up to 200 pc indicates a quasi-continuous injection of protons into the central molecular zone from a centrally located accelerator on a timescale $\Delta t$ exceeding the characteristic time of diffusive escape of particles from the central molecular zone, that is, $\Delta t \geq t_{\rm diff} \approx R^2/6D \approx 2\times 10^3 (D/10^{30})^{-1}$ yr, where $D$ (in cm$^2$ s$^{-1}$) is normalized to the characteristic value of multi-TeV cosmic rays in the Galactic disk[10]. In this regime the average injection rate of particles is found to be $\dot{Q}_p(\geq 10\ {\rm TeV}) \approx 4\times 10^{37}(D/10^{30})$ erg s$^{-1}$. The diffusion coefficient itself depends on the power spectrum of the turbulent magnetic field, which is unknown in the central molecular zone region. This introduces an uncertainty in the estimates of the injection power of relativistic protons. Yet, the diffusive nature of the propagation is constrained by the condition $R^2/6D \gg R/c$. For a radius of the central molecular zone region of 200 pc, this implies $D \ll 3\times 10^{30}$ cm$^2$ s$^{-1}$, and, consequently, $\dot{Q}_p \ll 1.2\times 10^{38}$ erg s$^{-1}$.

The energy spectrum of the diffuse $\gamma$-ray emission (Fig. 3) has been extracted from an annulus centred at Sagittarius (Sgr) A* (see Fig. 1). The best fit to the data is found for a spectrum following a power law extending with a photon index of ~2.3 to energies up to tens of TeV, without a cut-off or a break. This is the first time, to our knowledge, that such a $\gamma$-ray spectrum, arising from hadronic interactions, has been detected. Since these $\gamma$-rays result from the decay of neutral pions produced by $pp$ interactions, the derivation of such a hard power-law spectrum implies that the spectrum of the parent protons should extend to energies close to 1 PeV. The best fit of a $\gamma$-ray spectrum from neutral pion decay to the HESS data is found for a proton spectrum following a pure power law with an index of ~2.4. We note that $pp$ interactions of 1 PeV protons could also be studied by the observation of emitted neutrinos or X-rays from the synchrotron emission of secondary electrons and positrons (see Methods and Extended Data Figs 3 and 4). However, the measured $\gamma$-ray flux puts the expected fluxes of neutrinos and X-rays below or at best close to the sensitivities of the current instruments. Assuming a cut-off in the parent proton spectrum, the corresponding secondary $\gamma$-ray spectrum deviates from the HESS data at 68%, 90% and 95% confidence levels for cut-offs at 2.9 PeV, 0.6 PeV and 0.4 PeV, respectively. This is the first robust detection of a VHE cosmic hadronic accelerator which operates as a source of PeV particles (a 'PeVatron').

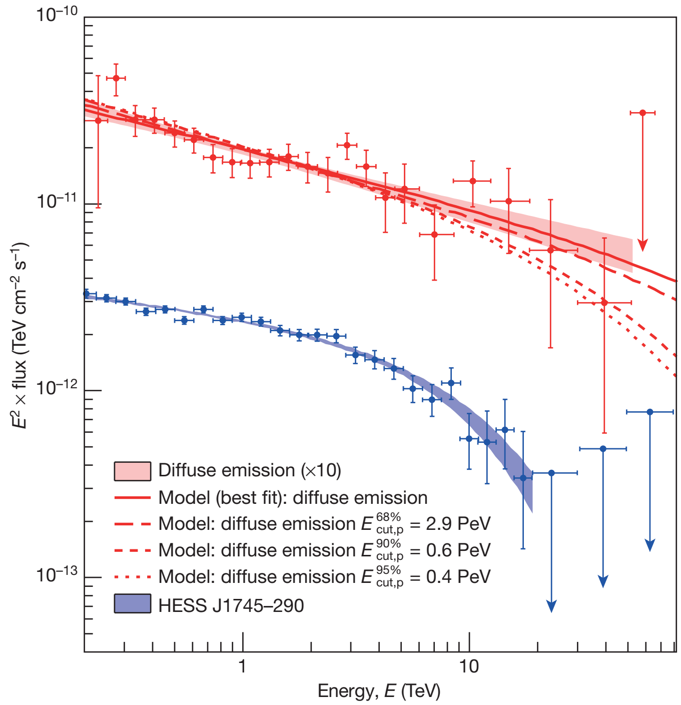


**Figure 3 | VHE $\gamma$-ray spectra of the diffuse emission and HESS J1745−290.** The $y$ axis shows fluxes multiplied by a factor $E^2$, where $E$ is the energy on the $x$ axis, in units of TeV cm$^{-2}$ s$^{-1}$. The vertical and horizontal error bars show the 1$\sigma$ statistical error and the bin size, respectively. Arrows represent 2$\sigma$ flux upper limits. The 1$\sigma$ confidence bands of the best-fit spectra of the diffuse and HESS J1745−290 are shown in red and blue shaded areas, respectively. Spectral parameters are given in Methods. The red lines show the numerical computations assuming that $\gamma$-rays result from the decay of neutral pions produced by proton–proton interactions. The fluxes of the diffuse emission spectrum and models are multiplied by 10 to visually separate them from the HESS J1745−290 spectrum.

Remarkably, the Galactic Centre PeVatron appears to be located in the same region as the central $\gamma$-ray source HESS J1745−290 (refs 11–14). Unfortunately, the current data cannot provide an answer as to whether there is an intrinsic link between these two objects. The point-like source HESS J1745−290 itself remains unidentified. Besides Sgr A* (ref. 15), other potential counterparts are the pulsar wind nebula G 359.95−0.04 (refs 16, 17) and a spike of annihilating dark matter[18]. Moreover, it has also been suggested that this source might have a diffuse origin, peaking towards the direction of the Galactic Centre because of the higher concentration there of both gas and relativistic particles[15]. In fact, this interpretation would imply an extension of the spectrum of the central source to energies beyond 10 TeV, which however is at odds with the detection of a clear cut-off in the spectrum of HESS J1745−290 at about 10 TeV (refs 19, 20; Fig. 3). Yet the attractive idea of explaining the entire $\gamma$-ray emission from the Galactic Centre by run-away protons from the same centrally located accelerator can still be compatible with the cut-off in the spectrum of the central source. For example, the cut-off could be due to the absorption of $\gamma$-rays caused by interactions with the ambient infrared radiation field. It should be noted that although the question of the link between the central $\gamma$-ray source and the proton PeVatron is interesting, it does not have a direct impact on the main conclusions of the present work.

The integration of the cosmic-ray radial distribution (Fig. 2) yields the total energy $W_p$ of $E \geq 10$ TeV protons confined in the central molecular zone: $W_p \approx 1.0 \times 10^{49}$ erg. A single supernova remnant (SNR) would suffice to provide this rather modest energy in cosmic rays. A possible candidate could be Sgr A East. Although this object has already been excluded as a counterpart of HESS J1745−290 (ref. 21), the multi-TeV protons accelerated by this object and then injected into the central molecular zone could contribute to the diffuse $\gamma$-ray component. Other potential sites of the acceleration of protons in the Galactic Centre are the compact stellar clusters[22]. Formally, the mechanical power in these clusters in the form of stellar winds, which can provide adequate conditions for particle acceleration, is sufficient to explain the required total energy of cosmic rays in the central molecular zone. However, the acceleration of protons to PeV energies requires bulk motions in excess of 10,000 km s$^{-1}$ which could only exist in the stellar clusters because of very young supernova shocks[23]. Thus, the mechanism of operation of PeVatrons in stellar clusters is reduced to the presence of supernova shocks. On the other hand, since the acceleration of PeV particles by shocks, either in individual SNRs or in stellar clusters, cannot last much longer than 100 years (ref. 24), we would need more than 10 supernova events to meet the requirement of continuous injection of cosmic rays in the central molecular zone over $\gg 10^3$ years. For the central 10 pc region, such a high supernova rate seems unlikely.

We suggest that the supermassive black hole at the Galactic Centre (Sgr A*) is the most plausible supplier of ultra-relativistic protons and nuclei; these particles could have been accelerated either in the accretion flow (that is, in the immediate vicinity of the black hole[15,25]) or somewhat further away—for example, at the site of termination of an outflow[26]. If Sgr A* is indeed the particles' source, the required acceleration rate of about $10^{37}$–$10^{38}$ erg s$^{-1}$ would exceed by two or three orders of magnitude the current bolometric luminosity of this object[27], and would constitute at least 1% of the current power produced by accretion onto the supermassive black hole. Given that the current accretion rate is relatively modest, and that at certain epochs this supermassive ($4 \times 10^6$ solar masses) black hole could have operated at a much higher accretion rate, we speculate that this higher rate could also facilitate greater cosmic-ray production rates[25]. We note that an average acceleration rate of $10^{39}$ erg s$^{-1}$ of $E > 10$ TeV protons over the last $10^6$–$10^7$ years would be sufficient to explain the flux of cosmic rays around the energy spectrum feature—the so-called 'knee'—at 1 PeV. If this explanation is correct, it could be a solution to one of the most controversial and actively debated problems of the paradigm of the SNR origin of Galactic cosmic rays[24,28,29].

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 2 April 2015; accepted 4 January 2016.**
**Published online 16 March 2016.**

1. Berezinskii, V. S., Bulanov, S. V., Dogiel, V. A. & Ptuskin, V. S. in *Astrophysics of Cosmic Rays* (ed. Ginzburg, V. L.) 33–38 (North-Holland, 1990).
2. Malkov, M. A. & Drury, L. O. Nonlinear theory of diffusive acceleration of particles by shock waves. *Rep. Prog. Phys.* **64,** 429–481 (2001).
3. Hillas, A. M. Evolution of ground-based gamma-ray astronomy from the early days to the Cherenkov Telescope Arrays. *Astropart. Phys.* **43,** 19–43 (2013).
4. Aharonian, F. A. Gamma rays from supernova remnants. *Astropart. Phys.* **43,** 71–80 (2013).
5. Clavel, M. *et al.* Echoes of multiple outbursts of Sagittarius A* revealed by Chandra. *Astron. Astrophys.* **558,** A32 (2013).
6. Su, M., Slatyer, T. R. & Finkbeiner, D. P. Giant gamma-ray bubbles from Fermi-LAT: active galactic nucleus activity or bipolar galactic wind? *Astrophys. J.* **724,** 1044–1082 (2010).
7. Aharonian, F. *et al.* Discovery of very-high-energy $\gamma$-rays from the Galactic Centre ridge. *Nature* **439,** 695–698 (2006).
8. Yang, R., Jones, D. I. & Aharonian, F. Fermi-LAT observations of the Sagittarius B complex. *Astron. Astrophys.* **580,** A90 (2015).
9. Aharonian, F. *Very High Energy Cosmic Gamma Radiation: A Crucial Window on the Extreme Universe* (World Scientific, 2004).
10. Strong, A. W., Moskalenko, I. V. & Ptuskin, V. S. Cosmic-ray propagation and interactions in the Galaxy. *Annu. Rev. Nucl. Part. Sci.* **57,** 285–327 (2007).
11. Aharonian, F. *et al.* Very high-energy gamma rays from the direction of Sagittarius A*. *Astron. Astrophys.* **425,** L13–L17 (2004).
12. Kosack, K. *et al.* TeV gamma-ray observations of the Galactic Center. *Astrophys. J.* **608,** L97–L100 (2004).
13. Tsuchiya, K. *et al.* Detection of sub-TeV gamma rays from the Galactic center direction by CANGAROO-II. *Astrophys. J.* **606,** L115–L118 (2004).
14. Albert, J. *et al.* Observation of gamma rays from the Galactic center with the MAGIC telescope. *Astrophys. J.* **638,** L101–L104 (2006).
15. Aharonian, F. & Neronov, A. High-energy gamma rays from the massive black hole in the Galactic center. *Astrophys. J.* **619,** 306–313 (2005).
16. Wang, Q. D., Lu, F. J. & Gotthelf, E. V. G359.95–0.04: pulsar candidate near Sgr A*. *Mon. Not. R. Astron. Soc.* **367,** 937–944 (2006).
17. Hinton, J. A. & Aharonian, F. Inverse Compton scenarios for the TeV gamma-ray emission of the Galactic centre. *Astrophys. J.* **657,** 302–307 (2007).
18. Belikov, A. V., Zaharijas, G. & Silk, J. Study of the gamma-ray spectrum from the Galactic Center in view of multi-TeV dark matter candidates. *Phys. Rev. D* **86,** 083516 (2012).
19. Aharonian, F. *et al.* Spectrum and variability of the Galactic center VHE $\gamma$-ray source HESS J1745–290. *Astron. Astrophys.* **503,** 817–825 (2009).
20. Archer, A. *et al.* Very-high energy observations of the Galactic center region by VERITAS in 2010–2012. *Astrophys. J.* **790,** 149 (2014).
21. HESS Collaboration. Localising the VHE $\gamma$-ray source at the Galactic Centre. *Mon. Not. R. Astron. Soc.* **402,** 1877–1882 (2010).
22. Crocker, R. M. *et al.* $\gamma$-rays and the far-infrared-radio continuum correlation reveal a powerful Galactic Centre wind. *Mon. Not. R. Astron. Soc.* **411,** L11–L15 (2011).
23. Bykov, A. M. Nonthermal particles and photons in starburst regions and superbubbles. *Astron. Astrophys. Rev.* **22,** 1–54 (2014).
24. Bell, A., Schure, K., Reville, B. & Giacinti, G. Cosmic ray acceleration and escape from supernova remnants. *Mon. Not. R. Astron. Soc.* **431,** 415–429 (2013).
25. Istomin, Y. N. On the origin of galactic cosmic rays. *New Astron.* **27,** 13–18 (2014).
26. Atoyan, A. & Dermer, C. D. TeV emission from the Galactic center black hole plerion. *Astrophys. J.* **617,** L123–L126 (2004).
27. Genzel, R., Eisenhauer, F. & Gillessen, S. The Galactic Center massive black hole and nuclear star cluster. *Rev. Mod. Phys.* **82,** 3121–3195 (2010).
28. Cristofari, P., Gabici, S., Casanova, S., Terrier, R. & Parizot, E. Acceleration of cosmic rays and gamma-ray emission from supernova remnants in the Galaxy. *Mon. Not. R. Astron. Soc.* **434,** 2748–2760 (2013).
29. Parizot, E. Cosmic ray origin: lessons from ultra-high-energy cosmic rays and the Galactic/extragalactic transition. *Nucl. Phys. B* **256–257** (Suppl.), 197–212 (2014).
30. Tsuboi, M., Handa, T. & Ukita, N. Dense molecular clouds in the Galactic Center region. I. Observations and data. *Astrophys. J.* **120** (Suppl.), 1–39 (1999).

**Acknowledgements** The support of the Namibian authorities and of the University of Namibia in facilitating the construction and operation of HESS is gratefully acknowledged, as is the support by the German Ministry for Education and Research (BMBF), the Max Planck Society, the German Research Foundation (DFG), the French Ministry for Research, the CNRS-IN2P3 and the Astroparticle Interdisciplinary Programme of the CNRS, the UK Science and Technology Facilities Council (STFC), the IPNP of the Charles University, the Czech Science Foundation, the Polish Ministry of Science and Higher Education, the South African Department of Science and Technology and National Research Foundation, and the University of Namibia. We thank the technical support staff in Berlin, Durham, Hamburg, Heidelberg, Palaiseau, Paris, Saclay and Namibia for the construction and operation of the equipment. R.C.G.C. is funded by an EU FP7 Marie Curie grant (PIEF-GA-2012-332350), J. Conrad is a Wallenberg Academy Fellow, and F.R. is a Heisenberg Fellow (DFG).

**Author Contributions** F.A., S.G., E.M. and A.V. analysed and interpreted the data, and prepared the manuscript. The whole HESS collaboration contributed to the publication, with involvement at various stages ranging from the design, construction and operation of the instrument, to the development and maintenance of all software for data handling, data reduction and data analysis. All authors reviewed, discussed and commented on the present results and on the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to HESS Collaboration (contact.hess@hess-experiment.eu).

**HESS Collaboration**
A. Abramowski[1], F. Aharonian[2,3,4], F. Ait Benkhali[2], A. G. Akhperjanian[4,5], E. O. Angüner[6], M. Backes[7], A. Balzer[8], Y. Becherini[9], J. Becker Tjus[10], D. Berge[11], S. Bernhard[12], K. Bernlöhr[2], E. Birsin[6], R. Blackwell[13], M. Böttcher[14], C. Boisson[15], J. Bolmont[16], P. Bordas[2], J. Bregeon[17], F. Brun[18], P. Brun[18], M. Bryan[8], T. Bulik[19], J. Carr[20], S. Casanova[2,21], N. Chakraborty[2], R. Chalme-Calvet[16], R. C. G. Chaves[17], A. Chen[22], M. Chrétien[16], S. Colafrancesco[22], G. Cologna[23], J. Conrad[24], C. Couturier[16], Y. Cui[25], I. D. Davids[7,14], B. Degrange[26], C. Deil[2], P. deWilt[13], A. Djannati-Ataï[27], W. Domainko[2], A. Donath[2], L. O'C. Drury[3], G. Dubus[28], K. Dutson[29], J. Dyks[30], M. Dyrda[21], T. Edwards[2], K. Egberts[31], P. Eger[2], J.-P. Ernenwein[20], P. Espigat[27], C. Farnier[24], S. Fegan[26], F. Feinstein[17], M. V. Fernandes[1], D. Fernandez[17], A. Fiasson[32], G. Fontaine[26], A. Förster[2], M. Füßling[33], S. Gabici[27], M. Gajdus[6], Y. A. Gallant[17], T. Garrigoux[16], G. Giavitto[33], B. Giebels[26], J. F. Glicenstein[18], D. Gottschall[25], A. Goyal[34], M.-H. Grondin[35], M. Grudzińska[19], D. Hadasch[12], S. Häffner[36], J. Hahn[2], J. Hawkes[13], G. Heinzelmann[1], G. Henri[28], G. Hermann[2], O. Hervet[15], A. Hillert[2], J. A. Hinton[2,29], W. Hofmann[2], P. Hofverberg[2], C. Hoischen[31], M. Holler[26], D. Horns[1], A. Ivascenko[14], A. Jacholkowska[16], M. Jamrozy[34], M. Janiak[30], F. Jankowsky[23], I. Jung-Richardt[36], M. A. Kastendieck[1], K. Katarzyński[37], U. Katz[36], D. Kerszberg[16], B. Khélifi[27], M. Kieffer[16], S. Klepser[33], D. Klochkov[25], W. Kluźniak[30], D. Kolitzus[12], Nu. Komin[22], K. Kosack[18], S. Krakau[10], F. Krayzel[32], P. P. Krüger[14], H. Laffon[35], G. Lamanna[32], J. Lau[13], J. Lefaucheur[27], V. Lefranc[18], A. Lemiére[27], M. Lemoine-Goumard[35], J.-P. Lenain[16], T. Lohse[6], A. Lopatin[36], C.-C. Lu[2], R. Lui[2], V. Marandon[2], A. Marcowith[17], C. Mariaud[26], R. Marx[2], G. Maurin[32], N. Maxted[17], M. Mayer[6], P. J. Meintjes[38], U. Menzler[10], M. Meyer[24], A. M. W. Mitchell[2], R. Moderski[30], M. Mohamed[23], K. Morå[24], E. Moulin[18], T. Murach[6], M. de Naurois[26], J. Niemiec[21], L. Oakes[6], H. Odaka[2], S. Öttl[12], S. Ohm[33], B. Opitz[1], M. Ostrowski[34], I. Oya[33], M. Panter[2], R. D. Parsons[2], M. Paz Arribas[6], N. W. Pekeur[14], G. Pelletier[28], P.-O. Petrucci[28], B. Peyaud[18], S. Pita[27], H. Poon[2], H. Prokoph[9], G. Pühlhofer[25], M. Punch[27], A. Quirrenbach[23], S. Raab[36], I. Reichardt[27], A. Reimer[12], O. Reimer[12], M. Renaud[17], R. de los Reyes[2], F. Rieger[2,39], C. Romoli[3], S. Rosier-Lees[32], G. Rowell[13], B. Rudak[30], C. B. Rulten[15], V. Sahakian[4,5], D. Salek[40], D. A. Sanchez[32], A. Santangelo[25], M. Sasaki[25], R. Schlickeiser[10], F. Schüssler[18], A. Schulz[33], U. Schwanke[6], S. Schwemmer[23], A. S. Seyffert[14], R. Simoni[8], H. Sol[15], F. Spanier[14], G. Spengler[24], F. Spies[1], Ł. Stawarz[34], R. Steenkamp[7], C. Stegmann[31,33], F. Stinzing[36], K. Stycz[33], I. Sushch[14], J.-P. Tavernet[16], T. Tavernier[27], A. M. Taylor[3], R. Terrier[27], M. Tluczykont[1], C. Trichard[32], R. Tuffs[2], K. Valerius[36], J. van der Walt[14], C. van Eldik[36], B. van Soelen[38], G. Vasileiadis[17], J. Veh[36], C. Venter[14], A. Viana[2], P. Vincent[16], J. Vink[8], F. Voisin[13], H. J. Völk[2], T. Vuillaume[28], S. J. Wagner[23], P. Wagner[6], R. M. Wagner[24], M. Weidinger[10], Q. Weitzel[2], R. White[29], A. Wierzcholska[21,23], P. Willmann[36], A. Wörnlein[36], D. Wouters[18], R. Yang[2], V. Zabalza[29], D. Zaborov[26], M. Zacharias[23], A. A. Zdziarski[30], A. Zech[15], F. Zefi[26] & N. Żywucka[34]

[1] Universität Hamburg, Institut für Experimentalphysik, Luruper Chaussee 149, D 22761 Hamburg, Germany. [2] Max-Planck-Institut für Kernphysik, PO Box 103980, D 69029 Heidelberg, Germany. [3] Dublin Institute for Advanced Studies, 31 Fitzwilliam Place, Dublin 2, Ireland. [4] National Academy of Sciences of the Republic of Armenia, Marshall Baghramian Avenue, 24, 0019 Yerevan, Armenia. [5] Yerevan Physics Institute, 2 Alikhanian Brothers Street, 375036 Yerevan, Armenia. [6] Institut für Physik, Humboldt-Universität zu Berlin, Newtonstrasse 15, D 12489 Berlin, Germany. [7] University of Namibia, Department of Physics, Private Bag 13301, Windhoek, Namibia. [8] GRAPPA, Anton Pannekoek Institute for Astronomy, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands. [9] Department of Physics and Electrical Engineering, Linnaeus University, 351 95 Växjö, Sweden. [10] Institut für Theoretische Physik, Lehrstuhl IV: Weltraum und Astrophysik, Ruhr-Universität Bochum, D 44780 Bochum, Germany. [11] GRAPPA, Anton Pannekoek Institute for Astronomy and Institute of High-Energy Physics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands. [12] Institut für Astro- und Teilchenphysik, Leopold-Franzens-Universität Innsbruck, A-6020 Innsbruck, Austria. [13] School of Chemistry and Physics, University of Adelaide, Adelaide 5005, Australia. [14] Centre for Space Research, North-West University, Potchefstroom 2520, South Africa. [15] LUTH, Observatoire de Paris, CNRS, Université Paris Diderot, 5 Place Jules Janssen, 92190 Meudon, France. [16] LPNHE, Université Pierre et Marie Curie Paris 6, Université Denis Diderot Paris 7, CNRS/IN2P3, 4 Place Jussieu, F-75252, Paris Cedex 5, France. [17] Laboratoire Univers et Particules de Montpellier, Université Montpellier 2, CNRS/IN2P3, CC 72, Place Eugène Bataillon, F-34095 Montpellier Cedex 5, France. [18] DSM/Irfu, CEA Saclay, F-91191 Gif-Sur-Yvette Cedex, France. [19] Astronomical Observatory, The University of Warsaw, Al. Ujazdowskie 4, 00-478 Warsaw, Poland. [20] Aix Marseille Université, CNRS/IN2P3, CPPM UMR 7346, 13288 Marseille, France. [21] Instytut Fizyki Jądrowej PAN, ul. Radzikowskiego 152, 31-342 Kraków, Poland. [22] School of Physics, University of the Witwatersrand, 1 Jan Smuts Avenue, Braamfontein, Johannesburg 2050, South Africa. [23] Landessternwarte, Universität Heidelberg, Königstuhl, D 69117 Heidelberg, Germany. [24] Oskar Klein Centre, Department of Physics, Stockholm University, Albanova University Center, SE-10691 Stockholm, Sweden. [25] Institut für Astronomie und Astrophysik, Universität Tübingen, Sand 1, D 72076 Tübingen, Germany. [26] Laboratoire Leprince-Ringuet, Ecole Polytechnique, CNRS/IN2P3, F-91128 Palaiseau, France. [27] APC, AstroParticule et Cosmologie, Université Paris Diderot, CNRS/IN2P3, CEA/Irfu, Observatoire de Paris, Sorbonne Paris Cité, 10, rue Alice Domon et Léonie Duquet, 75205 Paris Cedex 13, France. [28] Université Grenoble Alpes, IPAG, F-38000 Grenoble, France; CNRS, IPAG, F-38000 Grenoble, France. [29] Department of Physics and Astronomy, The University of Leicester, University Road, Leicester LE1 7RH, UK. [30] Nicolaus Copernicus Astronomical Center, ul. Bartycka 18, 00-716 Warsaw, Poland. [31] Institut für Physik und Astronomie, Universität Potsdam, Karl-Liebknecht-Strasse 24/25, D 14476 Potsdam, Germany. [32] Laboratoire d'Annecy-le-Vieux de Physique des Particules, Université Savoie Mont-Blanc, CNRS/IN2P3, F-74941 Annecy-le-Vieux, France. [33] DESY, D-15738 Zeuthen, Germany. [34] Obserwatorium Astronomiczne, Uniwersytet Jagielloński, ul. Orla 171, 30-244 Kraków, Poland. [35] Université Bordeaux, CNRS/IN2P3, Centre d'Études Nucléaires de Bordeaux Gradignan, 33175 Gradignan, France. [36] Universität Erlangen-Nürnberg, Physikalisches Institut, Erwin-Rommel-Strasse 1, D 91058 Erlangen, Germany. [37] Centre for Astronomy, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, Grudziadzka 5, 87-100 Torun, Poland. [38] Department of Physics, University of the Free State, PO Box 339, Bloemfontein 9300, South Africa. [39] ITA Universität Heidelberg, Germany. [40] GRAPPA, Institute of High-Energy Physics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.

## METHODS

**Data and spectral analyses.** The phase I of the HESS array consists of four identical imaging atmospheric Cherenkov telescopes. The study presented in the Letter makes use of data collected from 2004 to 2013. Data are taken in wobble mode; the pointing direction is chosen at an alternating offset of 0.7° to 1.1° from the target position[31]. Standard quality selection is applied to the data[31], and after the selection procedure the data set amounts to 226 h of live time at the nominal position of Sgr A*. The analysis technique used to select the $\gamma$-ray events is based on a semi-analytical model of the air shower development[32]. The background level is calculated in each position using the Ring Background method[33]. The map shown in Fig. 1 is the $\gamma$-ray excess count map per $0.02° \times 0.02°$ smoothed by the HESS point spread function and corrected for the telescope radial acceptance.

The spectral reconstruction is based on a forward-folding method[34], based itself on a maximum-likelihood procedure, comparing the energy distributions of signal and background events to predefined spectral shapes. The energy spectra (Fig. 3) are fitted by a power law, $\mathrm{d}N/\mathrm{d}E = \Phi_1 E^{-\Gamma_1}$, where $\Phi_1$ is the flux normalization and $\Gamma_1$ is the spectral index, and by a power law with an exponential cut-off, $\mathrm{d}N/\mathrm{d}E = \Phi_0 E^{-\Gamma_0} \times \exp(-E/E_{\mathrm{cut}})$, where $\Phi_0$ is the flux normalization, $\Gamma_0$ is the spectral index and $E_{\mathrm{cut}}$ is the cut-off energy; here $E$ is in TeV. A likelihood-ratio test between these two representations is performed to determine whether a significant deviation from a pure power law is preferred by the data. The best-fit spectra together with their $1\sigma$ confidence-level band are shown in Fig. 3 (shaded red and blue regions). The spectral points are uncorrelated flux points, obtained from the reconstructed best-fit spectrum. The error bars are $\pm 1\sigma$ (assuming a Poissonian distribution) of the excess number of events in the energy bin.

The diffuse emission spectrum is extracted from an annulus centred at Sgr A* (right panel of Fig. 1) with inner and outer radii of 0.15° and 0.45°, respectively, and a solid angle of $1.4 \times 10^{-4}$ sr. The best-fit spectrum is given by a power law with $\Phi_1 = (1.92 \pm 0.08_{\mathrm{stat}} \pm 0.28_{\mathrm{syst}}) \times 10^{-12}\,\mathrm{TeV}^{-1}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, and a photon index $\Gamma_1 = 2.32 \pm 0.05_{\mathrm{stat}} \pm 0.11_{\mathrm{syst}}$. Its $\gamma$-ray luminosity above 1 TeV is $L_\gamma(\geq 1\,\mathrm{Tev}) = (5.69 \pm 0.22_{\mathrm{stat}} \pm 0.85_{\mathrm{syst}}) \times 10^{34}\,\mathrm{erg}\,\mathrm{s}^{-1}$. The fit of a power law with an exponential energy cut-off is not preferred by the data. When compared with a pure power law, the likelihood-ratio test gives a $p$ value of 0.8 (~0.25 s.d. from the power-law fit). In order to investigate the possibility of spatial variations of the spectral indices over the central molecular zone, the spectra within all the regions (left panel of Fig. 1) are reconstructed. All the indices are compatible, within $1\sigma$, with an index of $2.32 \pm 0.05_{\mathrm{stat}} \pm 0.10_{\mathrm{syst}}$ (see Extended Data Table 3). The spectrum of the central source is extracted from a circular region of radius 0.1° centred on Sgr A*. The best-fit spectrum is a power law with an exponential cut-off with $\Phi_0 = (2.55 \pm 0.04_{\mathrm{stat}} \pm 0.37_{\mathrm{syst}}) \times 10^{-12}\,\mathrm{TeV}^{-1}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$, a photon index of $\Gamma_0 = 2.14 \pm 0.02_{\mathrm{stat}} \pm 0.10_{\mathrm{syst}}$, and an energy cut-off at $E_{\mathrm{cut}} = (10.7 \pm 2.0_{\mathrm{stat}} \pm 2.1_{\mathrm{syst}})$ TeV. When compared with a pure power law, the likelihood-ratio test gives a $p$ value of $3 \times 10^{-5}$. A power law with an exponential cut-off is clearly preferred by the data.

The diffuse $\gamma$-ray spectrum from the decay of neutral pions produced by $pp$ interactions reflects the parent proton spectrum. The normalization of the proton spectrum depends on a combination of the injection power, target mass and propagation effects. Its shape, however, is completely defined by the observed $\gamma$-ray spectrum. In the case of a $\gamma$-ray spectrum following a power law with index $\Gamma_1$, the parent proton spectrum should follow a power law with an index $\Gamma_p \approx \Gamma_1 + 0.1$ (ref. 9). Using the parameterization of ref. 35, the best-fit proton spectrum to the HESS data is obtained for a power law with $\Gamma_p \approx 2.4$. The corresponding $1\sigma$ confidence band of this $\gamma$-ray spectrum is described by the red shaded area in Fig. 3. A fit to the data is also done with the $\gamma$-ray spectral shape derived from a proton spectrum following a power law with an exponential cut-off. When compared with a power law, the likelihood-ratio test gives a $p$ value of 0.9 (~0.12 s.d. from the power-law fit), and it is thus not preferred by the data. The lower limits on the proton spectrum energy cut-off can thus be derived from this fit. The 68%, 90% and 95% confidence level deviation from the HESS data are found for proton spectra following a power law with $\Gamma_p \approx 2.4$ and cut-offs ($E^{68\%}_{\mathrm{cut,p}}$, $E^{90\%}_{\mathrm{cut,p}}$ and $E^{95\%}_{\mathrm{cut,p}}$) at 2.9 PeV, 0.6 PeV and 0.4 PeV, respectively. Their corresponding $\gamma$-ray spectra are plotted in Fig. 3 (red long-dashed, dashed and dotted lines, respectively). The numerical computation of the energy spectrum of cosmic-ray protons with energy >10 TeV escaping from the central source at a rate of $\sim 8 \times 10^{37}\,\mathrm{erg}\,\mathrm{s}^{-1}$ over a time of $\sim 6 \times 10^{13}$ yr is shown in Fig. 3 (red solid line). Their injection spectrum is $Q_p(E) \propto E^{-2.2}$ extending up to 4 PeV, and their transport is described by a diffusion coefficient $D(E) = 6 \times 10^{29} (E/10)^{\beta}\,\mathrm{cm}^2\,\mathrm{s}^{-1}$, with $\beta = 0.3$ ($E$ is in TeV). The resulting $\gamma$-ray spectrum is compatible with the best-fit diffuse spectrum.

**Mass estimation and cosmic-ray energy density in the central molecular zone.** The derivation of the cosmic-ray density profile in the central molecular zone rests on the distribution of target material (for cosmic-ray interactions). The bulk of the gas in the Galactic Centre region is in the form of the molecular hydrogen ($H_2$), which is very difficult to detect directly. Therefore indirect methods of estimating the mass using tracer molecules must be applied. Tracer molecules are typically rare relative to $H_2$ but much easier to detect and with an approximately known ratio to $H_2$. The mass estimates used in this Letter are based on the line emission of the CS molecule at the $J = 1$–0 transition[30]. In order to evaluate the systematic uncertainties in these estimates, other channels, such as the line emission from transitions of $^{12}C^{16}O$ (ref. 36) and HCN (ref. 37) molecules, are also invoked. The total mass in the inner 150 pc of the central molecular zone is estimated to be $(3^{+2}_{-1}) \times 10^7 M_\odot$ (refs 30, 38; $M_\odot$, solar mass) The regions shown in Fig. 1 (left) almost completely cover the inner 150 pc of the central molecular zone and are used to extract the radial distribution of cosmic rays. They are symmetrically distributed around the centre of the central molecular zone, which is offset from Sgr A* by ~50 pc in positive Galactic longitudes ($l \approx 0.33°$). These regions are described as follows.

*Ring 1 (R1):* semi-annulus with $[r_{\mathrm{in}}, r_{\mathrm{out}}] = [0.1°, 0.15°]$, where $r_{\mathrm{in}}$ and $r_{\mathrm{out}}$ are the inner and outer radii, respectively. A section of 66° is excluded in order to avoid a newly detected source which will be reported elsewhere. This section is bounded by the opening angles of +10° and −56° from the positive Galactic longitude axis (see Fig. 1). The average radial distance from Sgr A* is $r_d = 18.5$ pc.

*Ring 2 (R2):* semi-annulus $[r_{\mathrm{in}}, r_{\mathrm{out}}] = [0.15°, 0.2°]$, $r_d = 25.9$ pc.

*Ring 3 (R3):* semi-annulus $[r_{\mathrm{in}}, r_{\mathrm{out}}] = [0.2°, 0.3°]$, $r_d = 37.1$ pc.

*Circle 1/1b (C1/C1b):* circular region with 0.1° of radius centred at $l = 0.344°/359.544°$, $b = -0.4588°$, in Galactic coordinates, and $r_d = 59.3$ pc.

*Circle 2/2b (C2/C2b):* circular region with 0.1° of radius centred at $l = 0.544°/359.344°$, $b = -0.04588°$, $r_d = 89.0$ pc.

*Circle 3 (C3):* circular region with 0.1° of radius centred at $l = 0.744°$, $b = -0.04588°$, $r_d = 118.6$ pc.

*Circle 4 (C4):* circular region with 0.1° of radius centred at $l = 0.944°$, $b = -0.04588°$, $r_d = 148.3$ pc.

*Circle 5 (C5):* circular region with 0.1° of radius centred at $l = 1.144°$, $b = -0.04588°$, $r_d = 178.0$ pc.

If the $\gamma$-ray emission is completely due to the decay of neutral pions produced in proton–proton interactions, then the $\gamma$-ray luminosity $L_\gamma$ above energy $E_\gamma$ is related to the total energy of cosmic-ray protons $W_p$ as

$$L_\gamma(\geq E_\gamma) \approx \eta_N \frac{W_p(\geq 10E_\gamma)}{t_{pp \to \pi^0}} \tag{1}$$

where $t_{pp \to \pi^0} = 1.6 \times 10^8\,\mathrm{yr}\,(1\,\mathrm{cm}^{-3}/n_{\mathrm{H}})$ is the proton energy loss timescale due to neutral pion production in an environment of hydrogen gas of density $n_{\mathrm{H}}$ (ref. 9), and $\eta_N \approx 1.5$ accounts for the presence of nuclei heavier than hydrogen in both cosmic rays and interstellar matter. The energy density (in eV cm$^{-3}$) of cosmic rays, $w_{\mathrm{CR}}$, averaged along the line of sight is then:

$$w_{\mathrm{CR}}(\geq 10E_\gamma) = \frac{W_p(\geq 10E_\gamma)}{V} \approx 1.8 \times 10^{-2} \left(\frac{\eta_N}{1.5}\right)^{-1} \left(\frac{L_\gamma(\geq E_\gamma)}{10^{34}}\right) \left(\frac{M}{10^6 M_\odot}\right)^{-1} \tag{2}$$

where $M$ is the mass of the relevant region in solar masses, and $L_\gamma$ is in erg s$^{-1}$. The $\gamma$-ray luminosity above 1 TeV and the mass estimates (based on three tracers) for all regions are presented in Extended Data Table 1. The cosmic-ray energy densities in different regions, given in units of $10^{-3}$ eV cm$^{-3}$, which is the value of the local cosmic-ray energy density $w_0(\geq 10\,\mathrm{Tev})$ (as measured in the solar neighbourhood), are presented in Extended Data Table 2.

The uncertainty in the cosmic-ray energy density comes basically from the uncertainty in the mass estimates. The independent estimates from different tracers result in cosmic-ray enhancement factors in the inner regions of the central molecular zone ($r \leq 25$ pc from the Galactic Centre) as follows: $16^{+10}_{-5}$ (CS), $22^{+14}_{-7}$ (CO) and $24^{+16}_{-8}$ (HCN). The cosmic-ray radial distribution is also computed for all the different channels. The results (shown in Fig. 2 for CS line tracer) are in good agreement with the $1/r$ profile with $\chi^2/\mathrm{d.o.f.} = 11.8/9$, $\chi^2/\mathrm{d.o.f.} = 9.4/9$ and $\chi^2/\mathrm{d.o.f.} = 11.0/8$ for mass estimates based on the CS, CO and HCN tracers, respectively. At the same time, the data are in obvious conflict with profiles of the type $w(r) \propto 1/r^2$ ($\chi^2/\mathrm{d.o.f.} = 73.2/9$ (CS), $\chi^2/\mathrm{d.o.f.} = 78.0/9$ (CO) and $\chi^2/\mathrm{d.o.f.} = 57.9/8$ (HCN)), and of the type $w(r) \propto$ constant ($\chi^2/\mathrm{d.o.f.} = 61.2/9$ (CS), $\chi^2/\mathrm{d.o.f.} = 45.6/9$ (CO) and $\chi^2/\mathrm{d.o.f.} = 77.1/8$ (HCN)). Finally, when fitting a $1/r^\alpha$ profile to the data, the best fit is found for $\alpha$ equal to $1.10 \pm 0.12$ (CS), $0.97 \pm 0.13$ (CO) and $1.24 \pm 0.12$ (HCN), with $\chi^2/\mathrm{d.o.f.} = 11.1/8$, $\chi^2/\mathrm{d.o.f.} = 9.34/8$ and $\chi^2/\mathrm{d.o.f.} = 6.5/7$, respectively, which confirms the preference for an $1/r$ density profile to describe the data.

**Spectral analysis within the central molecular zone.** The findings of a PeVatron accelerating protons in a quasi-continuous regime, over a sufficiently long period of time to fill the whole central molecular zone, implies that the $\gamma$-ray energy spectral shape should be spatially independent over the central molecular zone. The available statistics in each of these regions prevent us from testing the existence of a cut-off beyond 10 TeV. The indices of all the regions are presented in Extended Data Table 3. All the indices are compatible within $1\sigma$ with an index of $2.32 \pm 0.05_{\mathrm{stat}} \pm 0.10_{\mathrm{syst}}$, as measured in the annulus in the right panel of Fig. 1, and are used to derive the properties of the Galactic Centre PeVatron. The compatibility of the measured spectral indices over the 200 pc of the central molecular zone provides an additional piece of evidence for the scenario proposed in this Letter.

**Multi-TeV $\gamma$-rays of leptonic origin?** For the diffuse $\gamma$-ray emission of Galactic Centre, we consider quite specific conditions which strongly constrain the possible scenarios of $\gamma$-ray production. Two major radiation mechanisms are related to interactions of ultrarelativistic protons and electrons, with the dense gas in the central molecular zone and with the ambient infrared radiation fields, respectively. To explain multi-TeV $\gamma$-rays, the maximum energy of protons and electrons need to be as large as $\sim$1 PeV and $\sim$100 TeV, respectively. Additionally, these particles should effectively propagate and fill the entire central molecular zone. Whereas in the case of the hadronic scenario one needs to postulate an existence of a PeVatron in the Galactic Centre, any 'leptonic' model of $\gamma$-ray production should address the following questions: (i) whether the accelerator could be sufficiently effective to boost the energy of electrons up to $\geq$100 TeV under the severe radiative losses in the Galactic Centre; (2) whether these electrons can escape the sites of their production and propagate over distances of tens of parsecs; and (3) whether they can explain the observed hard spectrum of multi-TeV $\gamma$-rays.

Acceleration of electrons to multi-100 TeV energies is more difficult than acceleration of protons because of severe synchrotron and inverse Compton (IC) losses. Formally, acceleration of electrons to energies beyond 100 TeV is possible in the so-called extreme accelerators, where the acceleration proceeds at the maximum possible rate allowed by classical electrodynamics, $t_{\mathrm{acc}} \approx R_{\mathrm{L}}/c \approx 0.4(E/100)B^{-1}$ yr, where $E$ is the energy in TeV, $B$ is the magnetic field in $\mu$G, $c$ is the velocity of light and $R_{\mathrm{L}}$ is the Larmor radius. Even so, the escape of such energetic electrons from the accelerator, and their propagation far enough (tens of parsecs) to fill the central molecular zone, can be realized only for rather unrealistically weak magnetic fields and fast diffusion. Indeed, the propagation time over a distance $R$ (in pc) and for a particle diffusion coefficient $D$ (in cm$^2$ s$^{-1}$) is equal to $t_{\mathrm{diff}} = R^2/6D \approx 2 \times 10^3(R/200)^2(D/10^{30})$ yr and, for typical interstellar conditions, is much longer than the synchrotron loss time of electrons with energy $E_{\mathrm{e}}$ (in TeV), $t_{\mathrm{synch}} \approx 10(B/100)^{-2}(E_{\mathrm{e}}/100)^{-1}$ yr (here $B$ is in $\mu$G).

The efficiency of a given $\gamma$-ray emitting process is determined by the cooling time of particles through that specific channel compared to the characteristic times of other (radiative and non-radiative) processes. The cooling times of relativistic electrons in the central molecular zone are shown in Extended Data Fig. 1. While bremsstrahlung and IC scattering result in $\gamma$-ray emission, the ionization and synchrotron losses reduce the efficiency of $\gamma$-ray production. Bremsstrahlung is an effective mechanism of $\gamma$-radiation at GeV energies. Above 100 GeV, IC cooling becomes more effective ($t_{\mathrm{IC}} < t_{\mathrm{br}}$; where $t_{\mathrm{IC}}$ and $t_{\mathrm{br}}$ are the electrons' cooling times through IC and bremsstrahlung, respectively), and strongly dominates over bremsstrahlung at energies above 10 TeV. This can be seen in Extended Data Fig. 2, where the results of calculations of the spectral energy distribution of broad-band emission of electrons are shown. The calculations are performed for an acceleration spectrum following a power law with an exponential cut-off at 100 TeV. Assuming that electrons are injected in a continuous regime, the steady-state spectrum of electrons is obtained by solving the kinetic equation which takes into account the energy losses of electrons due to ionization, bremsstrahlung, synchrotron radiation and IC scattering. At low energies, the losses due to the diffusive escape of electrons from the central molecular zone are more important. Although it has been shown that the magnetic field in the Galactic Centre should have a lower limit of $B = 50$ $\mu$G on 400 pc scales[39], here we assume a very low $B = 15$ $\mu$G. Even with such low magnetic field, it is seen that above 10 TeV the calculations do not match the observed fluxes. If we assume, say, gas density higher by an order of magnitude (for example, if $\gamma$-rays are produced mainly in dense cores of molecular clouds), then bremsstrahlung would dominate over the IC contribution, and the flux of $\gamma$-rays could be increased. However, for any reasonable magnetic field, the synchrotron losses above 10 TeV will dominate over bremsstrahlung. This will make the steady-state electron spectrum steeper with power-law index $\alpha = \alpha_0 + 1$ ($\alpha_0$ is the power-law index of the electron injection spectrum). Since the $\gamma$-ray spectrum produced owing to bremsstrahlung mimics the energy spectrum of parent electrons ($\Gamma = \alpha$), at energies of $\gamma$-rays above a few TeV we should expect quite a steep spectrum of $\gamma$-rays, with a power-law index $\Gamma > 3.4$. This is in apparent conflict with observations.

**Multiwavelength and multi-messenger signatures of PeVatrons.** Galactic PeVatrons have unique signatures which allow their unambiguous identification among other particle accelerators. Such signatures are related to the neutral secondary products resulting from hadronic interactions of accelerated PeV protons and nuclei ($E \geq 0.1$ PeV per nucleon) with the ambient gas. The secondaries produced at low energies, in particular MeV/GeV $\gamma$-rays and the radio synchrotron emission of primary and secondary electrons and positrons, do carry information about the accelerator, however, strictly speaking, they are not directly linked to the PeV particles. The extrapolations from low to high energies based on theoretical assumptions are model-dependent, and therefore biased. Obviously, they cannot substitute for direct measurements at highest energies.

All three products of interactions of ultrarelativistic protons ($\gamma$-rays, neutrinos and electrons) generated through the production and decay of $\pi^0$, $\pi^+$ and $\pi^-$ mesons, receive approximately 10% of the energy of primary protons, thus multi-TeV secondary neutrals carry unequivocal information about the primary PeV protons. First of all, this concerns $\geq$10 TeV $\gamma$-rays, because at such high energies the efficiency of leptonic channels of production of high energy $\gamma$-rays in general, and in the central molecular zone, in particular, is dramatically reduced (see above). The flux sensitivity as well as the angular and energy resolutions achieved by the HESS array allow adequate studies of the acceleration sites and the propagation of accelerated protons up to 1 PeV based on the morphological and spectral properties of multi-TeV $\gamma$-rays from the Galactic Centre.

A second independent and straightforward proof of the hadronic origin of diffuse $\gamma$-rays from the central molecular zone would be the detection of multi-TeV neutrinos spatially correlated with $\gamma$-rays. In Extended Data Fig. 3 we show the fluxes of high energy neutrinos which should accompany the $\gamma$-ray flux presented in Fig. 3. The calculations are based on the parent proton spectrum derived from $\gamma$-ray data, therefore the only free parameter in these calculations is the high energy cut-off $E_0$ in the spectrum of parent protons. The condition for the detection of high energy neutrinos by km$^3$-scale detectors (such as IceCube or KM3Net) can be expressed by a minimum flux of $\gamma$-rays, assuming that both neutrinos and $\gamma$-rays are products of $pp$ interactions. The estimate of detectability of neutrinos is most robust (that is, less sensitive to the spectral shape) when normalized at a particle energy of $\sim$20 TeV. Namely, neutrinos can be detected by a km$^3$ volume detector if the differential flux of accompanying $\gamma$-rays at 20 TeV exceeds $10^{-12}$ TeV cm$^{-2}$ s$^{-1}$ (ref. 40). Since the $\gamma$-ray fluxes (Fig. 3) are quite close to this value, we may conclude that multi-TeV neutrinos from the Galactic Centre can be marginally detected by a km$^3$-scale detector after several years of exposure.

The third complementary channel of information about the PeV protons is carried by the secondary electrons, through their synchrotron radiation. In $pp$ interactions, electrons are produced in a fair balance with neutrinos and photons (their distribution almost coincides with the spectrum of the electronic neutrinos)[35], that is, they carry a significant fraction of the energy of the incident proton. In environments with magnetic field $B \geq 100$ $\mu$G, the lifetime of secondary electrons producing X-rays of energy $\varepsilon_x$ (in keV) is quite short, $\sim 15(B/100)^{-3/2}(\varepsilon_x/10)^{-1/2}$ yr, compared to other characteristic times—for example, to the propagation time of protons over the central molecular zone. Therefore, hard X-rays can be considered as a 'prompt' radiation component emitted in hadronic interactions simultaneously with $\gamma$-rays and neutrinos. The energy release in X-rays calculated self-consistently for the proton spectrum derived from the $\gamma$-ray data exceeds 10% of the energy released in multi-TeV neutrinos and $\gamma$-rays (see in Extended Data Fig. 4). In general, this is quite substantial, given the superior flux sensitivity of X-ray instruments, especially for point like sources. However, since the radiation component shown in Extended Data Fig. 4 is integrated over a very large (for X-ray instruments) region with an angular size of $\sim$1°, it is overshadowed by the diffuse X-ray emission detected by XMM-Newton[41]. This makes the detection of this component practically impossible.

Finally, we mention relativistic neutrons as another potential messenger of hadronic processes produced in the Galactic Centre, predominantly in the reaction $pp \rightarrow pn\pi^+$. However, because of the short lifetime of $\tau \approx 10^3$ s, only neutrons of energy $E_n \geq m_n d/(c\tau) \approx 10^{18}$ eV (here $m_n$ is the neutron mass) could reach us before they decay during the "free flight" from the Galactic Centre ($d = 8.5$ kpc is the distance to the Galactic Centre). Thus if the proton spectrum in the Galactic Centre extends to extremely high energies, one can probe, in principle, this additional channel of information by detectors of cosmic rays, in particular by the AUGER observatory.

**Alternative sources of PeV cosmic rays in the Galactic Centre.** Several cosmic-ray sources are present in the Galactic Centre region. Besides the Galactic Centre supermassive black hole, discussed in this Letter, alternative sources of the cosmic rays responsible for the $\gamma$-ray emission from the central molecular zone include supernova remnants (SNRs)[42], stellar clusters[22], and radio filaments[43]. Any of these scenarios should satisfy the following conditions, derived from the HESS observations of the central molecular zone: (i) the accelerator has to be located in the inner $\sim$10 pc of the Galaxy, (ii) the accelerator(s) has(have) to be continuous over a timescale of at least thousands of years, and (iii) the acceleration has to proceed up to PeV energies.

**SNRs.** The acceleration mechanism operating at SNRs is widely believed to be diffusive shock acceleration, characterized by an acceleration timescale $t_{\rm acc} \approx D(E)/u_{\rm s}^2$, where $D(E) \propto E/B$ is the Bohm diffusion coefficient of a cosmic ray of energy $E$ in a magnetic field $B$, and $u_{\rm s}$ is the shock speed. It is clear, then, that the fastest acceleration rate is obtained for the largest possible values of the shock speed and of the magnetic field strength.

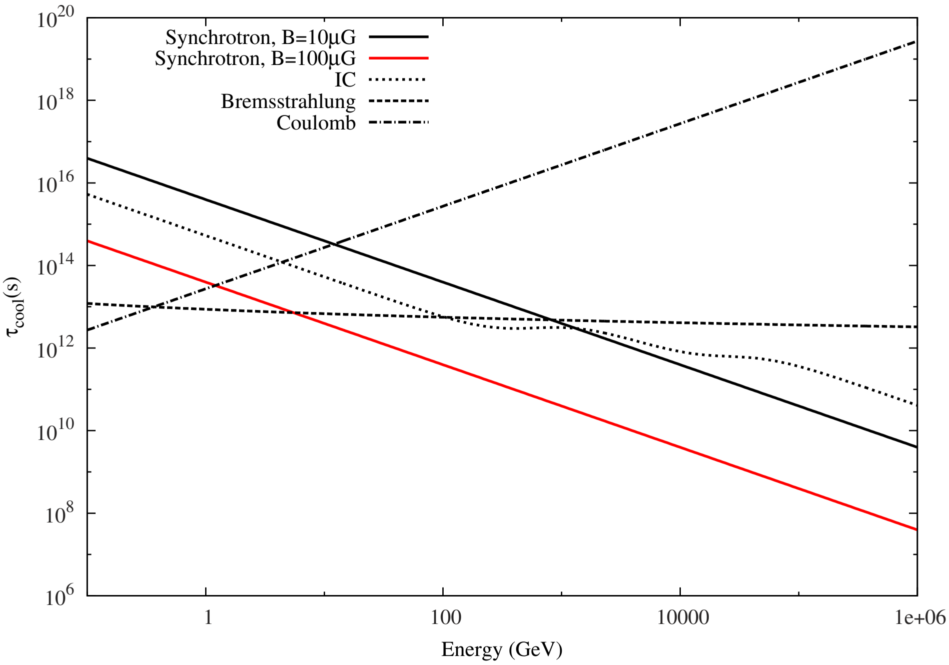
In the early free expansion phase of the SNR evolution, shock speeds as high as $\sim 10^4$ km s$^{-1}$ can be achieved. The magnetic field strength is also expected to be very large during this early phase (up to about 0.1–1 mG; ref. 44) due to efficient field amplification connected to the acceleration of cosmic rays at the shock[45]. Under these circumstances, SNRs can accelerate protons up to an energy of $E_{\rm max} \approx 10^{14}(B/100)(u_{\rm s}/10{,}000)^2(\Delta t_{\rm PeV})$ eV, where $\Delta t_{\rm PeV}$ (in yr) is the duration of the phase where shock speed and magnetic field are high enough to allow the acceleration of particles up to PeV energies ($B$ is in μG, $u_{\rm s}$ in km s$^{-1}$). It has been shown in ref. 24 that the duration of this phase is of the order of tens of years (definitely less than a century). Thus, even though SNRs can potentially provide PeV particles, they cannot act as (quasi) continuous injectors of such energetic particles for a time of the order of thousands of years.

**Stellar clusters.** Other possible cosmic-ray sources in the Galactic Centre region are stellar clusters. Three are known in the inner ~0.1°–0.2° region: the central, the Arches, and the Quintuplet cluster. The most likely sites for acceleration of particles in stellar clusters are the stellar winds of the massive OB stars that form the cluster, and the shocks of the supernovae which mark the end of the life of these stars. However, in order to accelerate particles up to PeV energies, very large shock velocities, of the order of at least 10,000 km s$^{-1}$ are needed[24]. Velocities of this order can hardly be found in stellar wind termination shocks, and thus SNR shocks following the explosion of cluster member stars remain the best candidates as particle accelerators.

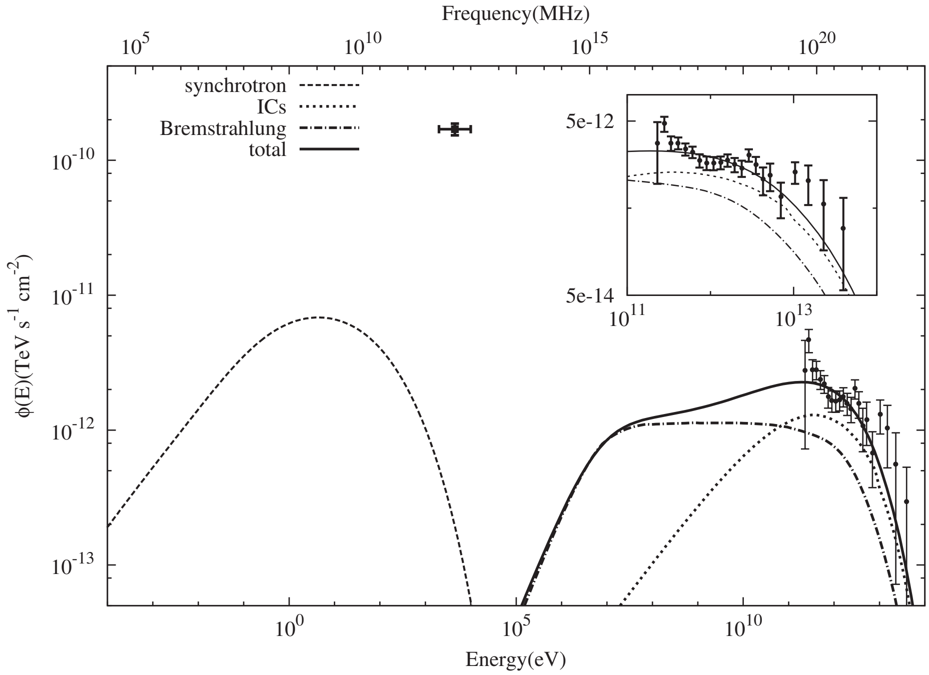
Both the Arches and the Quintuplet clusters are located outside the inner ~10 pc region[23], and this disfavours their role as accelerators of the cosmic rays responsible for the diffuse γ-ray emission. On the contrary, the central cluster is located well within the central 10 pc region, and thus should be considered as a potential candidate for the acceleration of cosmic rays in the central molecular zone. The γ-ray observations suggest that the cosmic-ray source in the Galactic Centre should act (quasi) continuously over time $t_{\rm inj}$ of a few thousands of years. Given that an SNR can accelerate PeV particles over a time interval $t_{\rm PeV}$ (in yr) of less than a century, we would need at least $\sim 10(t_{\rm Pev}/100)(t_{\rm inj}/1{,}000)^{-1}$ supernova explosions happening over the last $t_{\rm inj}$ (in yr) within the central cluster. Given the very small size of the region (~0.4 pc), such a large supernova explosion rate is unrealistic.

**Radio filaments.** It has been proposed in ref. 43 that the diffuse γ-ray emission from the central molecular zone was the result of non-thermal Bremsstrahlung from relativistic electrons[43,46]. In this scenario, the putative sources of γ-ray emitting electrons are the elongated radio filaments detected throughout the central molecular zone region[43]. This is in tension with one of the main finding of this Letter, that is, the location of the source of cosmic rays is in the inner ~10 pc of the Galaxy. Though the acceleration mechanism is not discussed in ref. 43, filaments are assumed to somehow accelerate electrons and then release them in the interstellar medium. In order to fill the whole central molecular zone region before being cooled by synchrotron and IC losses, electrons are assumed to propagate ballistically (that is, at the speed of light without a significant deflection in the magnetic field). This unconventional assumption is made at the expenses of a very large energy requirement: observations can be explained if the energy injection rate of cosmic-ray electrons in the central molecular zone is of the order of $\sim 10^{41}$ erg s$^{-1}$ (ref. 43). This is a very large injection rate, being comparable to the total luminosity of cosmic-ray protons in the whole Galaxy, and makes this scenario problematic.

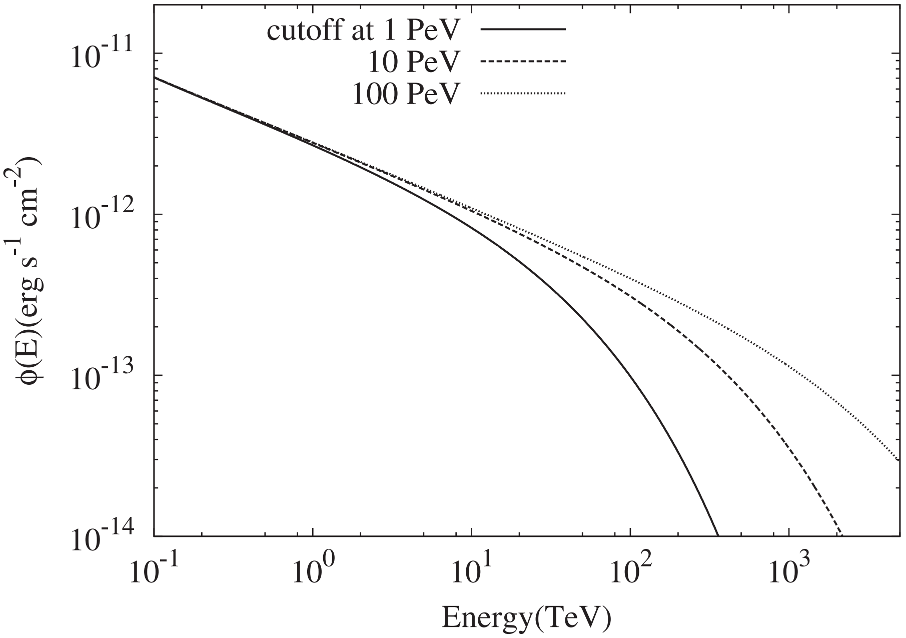
31. Aharonian, F. *et al.* Observations of the Crab Nebula with HESS. *Astron. Astrophys.* **457,** 899–915 (2006).
32. de Naurois, M. & Rolland, L. A high performance likelihood reconstruction of γ-rays for imaging atmospheric Cherenkov telescopes. *Astropart. Phys.* **32,** 231–252 (2009).
33. Berge, D., Funk, S. & Hinton, J. A. Background modelling in very-high-energy γ-ray astronomy. *Astron. Astrophys.* **466,** 1219–1229 (2007).
34. Piron, F. *et al.* Temporal and spectral gamma-ray properties of Mkn 421 above 250 GeV from CAT observations between 1996 and 2000. *Astron. Astrophys.* **374,** 895–906 (2001).
35. Kelner, S., Aharonian, F. & Bugayov, V. Energy spectra of gamma-rays, electrons and neutrinos produced at proton-proton interactions in the very high energy regime. *Phys. Rev. D* **74,** 034018 (2006); erratum **79,** 039901 (2009).
36. Oka, T. *et al.* A large-scale CO survey of the Galactic center. *Astrophys. J.* **118** (Suppl.), 455–515 (1998).
37. Jones, P. *et al.* Spectral imaging of the Central Molecular Zone in multiple 3-mm molecular lines. *Mon. Not. R. Astron. Soc.* **419,** 2961–2986 (2012).
38. Ferrière, K., Gillard, W. & Jean, P. Spatial distribution of interstellar gas in the innermost 3 kpc of our Galaxy. *Astron. Astrophys.* **467,** 611–627 (2007).
39. Crocker, R. M., Jones, D. I., Melia, F., Ott, J. & Protheroe, R. J. A lower limit of 50 microgauss for the magnetic field near the Galactic Centre. *Nature* **463,** 65–67 (2010).
40. Vissani, F., Aharonian, F. & Sahakyan, N. On the detectability of high-energy galactic neutrino sources. *Astropart. Phys.* **34,** 778–783 (2011).
41. Heard, V. & Warwick, R. S. XMM-Newton observations of the Galactic Centre region – I. The distribution of low-luminosity X-ray sources. *Mon. Not. R. Astron. Soc.* **428,** 3462–3477 (2013).
42. Büsching, I., de Jager, O. C. & Snyman, J. Obtaining cosmic-ray propagation parameters from diffuse very high energy gamma-ray emission from the galactic center ridge. *Astrophys. J.* **656,** 841–846 (2007).
43. Yusef-Zadeh, F. *et al.* Interacting cosmic rays with molecular clouds: a bremsstrahlung origin of diffuse high energy emission from the inner 2° × 1° of the Galactic center. *Astrophys. J.* **762,** 33 (2013).
44. Vink, J. Supernova remnants: the X-ray perspective. *Astron. Astrophys. Rev.* **20,** 49 (2012).
45. Bell, A. R. Turbulent amplification of magnetic field and diffusive shock acceleration of cosmic rays. *Mon. Not. R. Astron. Soc.* **353,** 550–558 (2004).
46. Yusef-Zadeh, F. *et al.* The origin of diffuse X-ray and γ-ray emission from the Galactic center region: cosmic ray particles. *Astrophys. J.* **656,** 847–869 (2007).
47. Strong, A. W. & Moskalenko, I. V. Propagation of cosmic-ray nucleons in the Galaxy. *Astrophys. J.* **509,** 212–228 (1998).



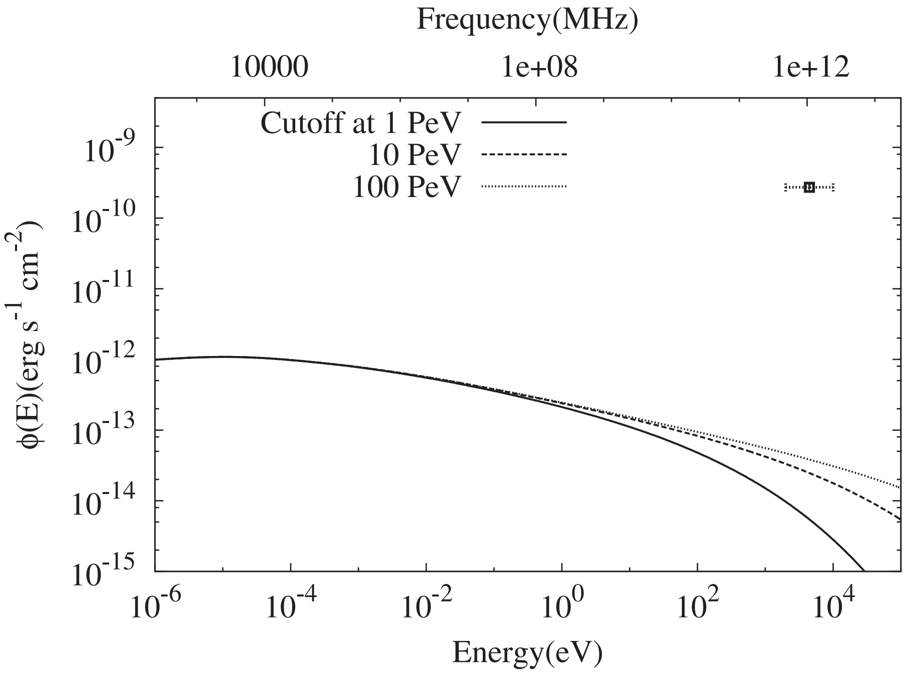
**Extended Data Figure 1 | Cooling times of electrons in the Galactic Centre as a function of energy.** The cooling times ($\tau_{\mathrm{cool}}$) due to ionization (or Coulomb) losses and bremsstrahlung are inversely proportional to the gas density $n$; here $n = 100\,\mathrm{cm}^{-3}$ is assumed. The cooling time of the synchrotron radiation is proportional to $1/B^2$, where $B$ is the magnetic field. The synchrotron cooling times are given for magnetic fields $B = 10\,\mu\mathrm{G}$ and $B = 100\,\mu\mathrm{G}$. The total energy densities of the cosmic microwave background and of local near-infrared (NIR) and far-infrared (FIR) infrared radiation fields used to calculated the cooling time due to the IC scattering are extracted from the GALPROP code[47]. The integrated densities are 17.0 eV cm$^3$ and 1.3 eV cm$^3$ for NIR and FIR, respectively.



**Extended Data Figure 2 | Broad-band spectral energy distribution of radiation by relativistic electrons.** The flux from synchrotron radiation, bremsstrahlung and IC scattering is compared to the fluxes of diffuse $\gamma$-ray emission measured by HESS (black points with vertical error bars). The flux of diffuse X-ray emission measured by XMM-Newton[41] (black point with horizontal error bar) and integrated over the central molecular zone region is also shown. Inset (top right) shows a zoomed view of the spectral energy distribution in the VHE range (100 GeV–100 TeV). The vertical and horizontal error bars show the $1\sigma$ statistical errors and the bin size, respectively.



**Extended Data Figure 3 | The spectral energy distribution of high energy neutrinos—the counterparts of diffuse γ-rays from the Galactic Centre.** The energy spectrum of parent protons is derived from the γ-ray data. The three curves correspond to different values of the exponential cut-off in the proton spectrum: 1 PeV, 10 PeV and 100 PeV.



**Extended Data Figure 4 | The spectral energy distribution of synchrotron radiation of secondary electrons produced in *pp* interactions.** The spectra of protons are the same as in Extended Data Fig. 3. The magnetic field is assumed to be 100 μG. The flux of diffuse X-ray emission measured by XMM-Newton and integrated over the central molecular zone region is also shown. The horizontal error bar corresponds to the bin size.

**Extended Data Table 1 | $\gamma$-ray luminosities and masses in different regions of the central molecular zone**

| Region | $L_\gamma(1\,\mathrm{TeV} \leq \mathrm{E}_\gamma)[\times 10^{34}\mathrm{erg\,s^{-1}}]$ | Mass $[\times 10^6 \mathrm{M}_\odot]$ | | |
|---|---|---|---|---|
| | | CS | CO | HCN |
| R1 | $1.54 \pm 0.06_{\mathrm{stat}} \pm 0.23_{\mathrm{sys}}$ | 1.35 | 0.88 | 0.88 |
| R2 | $1.21 \pm 0.05_{\mathrm{stat}} \pm 0.18_{\mathrm{sys}}$ | 1.11 | 0.98 | 0.86 |
| R3 | $2.05 \pm 0.08_{\mathrm{stat}} \pm 0.30_{\mathrm{sys}}$ | 2.65 | 2.01 | 1.65 |
| C1 | $0.73 \pm 0.03_{\mathrm{stat}} \pm 0.11_{\mathrm{sys}}$ | 1.51 | 1.00 | 1.02 |
| C1b | $0.63 \pm 0.03_{\mathrm{stat}} \pm 0.09_{\mathrm{sys}}$ | 1.49 | 1.03 | 0.82 |
| C2 | $0.84 \pm 0.03_{\mathrm{stat}} \pm 0.12_{\mathrm{sys}}$ | 1.84 | 1.40 | 1.34 |
| C2b | $0.51 \pm 0.02_{\mathrm{stat}} \pm 0.07_{\mathrm{sys}}$ | 1.01 | 0.86 | nc |
| C3 | $0.87 \pm 0.03_{\mathrm{stat}} \pm 0.13_{\mathrm{sys}}$ | 2.69 | 1.55 | 1.73 |
| C4 | $0.42 \pm 0.02_{\mathrm{stat}} \pm 0.06_{\mathrm{sys}}$ | 1.54 | 0.98 | 1.33 |
| C5 | $0.49 \pm 0.02_{\mathrm{stat}} \pm 0.07_{\mathrm{sys}}$ | 1.60 | 1.02 | 1.15 |

The errors quoted are $1\sigma$ statistical (stat) and systematic (sys) errors. The region C2b is not covered (nc) in the HCN line observations[37].

**Extended Data Table 2 | Cosmic-ray energy densities in different regions of the central molecular zone**

| Region | $w_{\mathrm{CR}}(\geq 10\mathrm{TeV})$ [$10^{-3}$ eV/cm$^3$] | | |
|---|---|---|---|
| | CS | CO | HCN |
| R1 | 17.3 ± 2.6 | 26.7 ± 4.1 | 26.6 ± 4.1 |
| R2 | 16.6 ± 2.5 | 18.6 ± 2.8 | 21.5 ± 3.3 |
| R3 | 11.8 ± 1.8 | 15.5 ± 2.4 | 18.8 ± 2.9 |
| C1 | 7.3 ± 1.1 | 11.0 ± 1.7 | 10.9 ± 1.6 |
| C1b | 6.5 ± 1.0 | 9.3 ± 1.4 | 11.6± 1.8 |
| C2 | 7.0 ± 1.0 | 9.1 ± 1.4 | 9.4 ± 1.4 |
| C2b | 7.7 ± 1.2 | 9.0 ± 1.4 | nc |
| C3 | 4.9 ± 0.7 | 8.5 ± 1.3 | 7.7 ± 1.1 |
| C4 | 4.1 ± 0.6 | 6.6 ± 0.9 | 4.8 ± 0.7 |
| C5 | 4.7 ± 0.7 | 7.4 ± 1.1 | 6.5 ± 1.0 |

The densities are given in units of $10^{-3}$ eV cm$^{-3}$, which is the value of the local cosmic-ray energy density measured in the solar neighbourhood. The errors quoted are $1\sigma$ statistical plus systematic errors. The region C2b is not covered (nc) in the HCN line observations[37].

**Extended Data Table 3 | Power-law spectral indices of the $\gamma$-ray energy spectrum in different regions of the central molecular zone**

| Region | Power-law spectral index |
|---|---|
| R1 | $2.27 \pm 0.05_{\mathrm{stat}} \pm 0.11_{\mathrm{sys}}$ |
| R2 | $2.33 \pm 0.07_{\mathrm{stat}} \pm 0.11_{\mathrm{sys}}$ |
| R3 | $2.18 \pm 0.07_{\mathrm{stat}} \pm 0.10_{\mathrm{sys}}$ |
| C1 | $2.38 \pm 0.10_{\mathrm{stat}} \pm 0.11_{\mathrm{sys}}$ |
| C1b | $2.26 \pm 0.10_{\mathrm{stat}} \pm 0.10_{\mathrm{sys}}$ |
| C2 | $2.39 \pm 0.08_{\mathrm{stat}} \pm 0.12_{\mathrm{sys}}$ |
| C2b | $2.44 \pm 0.14_{\mathrm{stat}} \pm 0.12_{\mathrm{sys}}$ |
| C3 | $2.35 \pm 0.06_{\mathrm{stat}} \pm 0.11_{\mathrm{sys}}$ |
| C4 | $2.23 \pm 0.10_{\mathrm{stat}} \pm 0.11_{\mathrm{sys}}$ |
| C5 | $2.27 \pm 0.11_{\mathrm{stat}} \pm 0.11_{\mathrm{sys}}$ |

The quoted errors are $1\sigma$ statistical (stat) and systematic (sys) errors.

doi:10.1038/nature17166

# Lunar true polar wander inferred from polar hydrogen

M. A. Siegler[1,2]*, R. S. Miller[3]*, J. T. Keane[4]*, M. Laneuville[5], D. A. Paige[6], I. Matsuyama[4], D. J. Lawrence[7], A. Crotts[8]‡ & M. J. Poston[9]

**The earliest dynamic and thermal history of the Moon is not well understood. The hydrogen content of deposits near the lunar poles may yield insight into this history, because these deposits (which are probably composed of water ice) survive only if they remain in permanent shadow. If the orientation of the Moon has changed, then the locations of the shadowed regions will also have changed. The polar hydrogen deposits have been mapped by orbiting neutron spectrometers[1–3], and their observed spatial distribution does not match the expected distribution of water ice inferred from present-day lunar temperatures[4,5]. This finding is in contrast to the distribution of volatiles observed in similar thermal environments at Mercury's poles[6]. Here we show that polar hydrogen preserves evidence that the spin axis of the Moon has shifted: the hydrogen deposits are antipodal and displaced equally from each pole along opposite longitudes. From the direction and magnitude of the inferred reorientation, and from analysis of the moments of inertia of the Moon, we hypothesize that this change in the spin axis, known as true polar wander, was caused by a low-density thermal anomaly beneath the Procellarum region. Radiogenic heating within this region resulted in the bulk of lunar mare volcanism[7–11] and altered the density structure of the Moon, changing its moments of inertia. This resulted in true polar wander consistent with the observed remnant polar hydrogen. This thermal anomaly still exists and, in part, controls the current orientation of the Moon. The Procellarum region was most geologically active early in lunar history[7–9], which implies that polar wander initiated billions of years ago and that a large portion of the measured polar hydrogen is ancient, recording early delivery of water to the inner Solar System. Our hypothesis provides an explanation for the antipodal distribution of lunar polar hydrogen, and connects polar volatiles to the geologic and geophysical evolution of the Moon and the bombardment history of the early Solar System.**

Lunar polar volatiles, including water ice, record the delivery, weathering and loss of external material, as well as the orbital dynamic history of the Moon[4,12,13]. Epithermal neutron deficits measured by orbital instruments provide an effective means of probing the spatial distribution and quantity of these volatiles through the measurement of hydrogen abundances. Here, we use improved data sets of lunar hydrogen abundance that are derived using a statistics-based likelihood analysis[14–16], shown in Fig. 1a, b. Our analysis relies on data from the Lunar Prospector Neutron Spectrometer (see Methods for a discussion of additional neutron data sets). Enhancements are determined on a pixel-by-pixel basis relative to the mid-latitude lunar highlands, which are assumed to be hydrogen-poor[14,16]. Modelled present-day temperature-dependent ice stability depths are also shown at the approximate spatial resolution of the Lunar Prospector Neutron Spectrometer (Fig. 1c, d).

These maps show four key features: first, the polar hydrogen maxima (north: 84.9° N, 147.9° E; south: 84.1° S, 309.4° E) are offset from the current rotation axis of the Moon by roughly 5.5°; second, the hydrogen enhancements are of similar magnitude at both poles; third, the asymmetric enhancements do not correlate with expectations from the current thermal or permanently shadowed environment[4,5,17]; and fourth, and most relevant to this study, the spatial distributions of polar hydrogen appear to be nearly antipodal.

A perfect antipodal relationship would manifest as identical distributions separated by 180° in longitude (Methods, Extended Data Fig. 1). This relationship can be quantified for any set of north–south spatial distributions by calculating $r(\alpha=0)$ and $r(\alpha)$, where $r(\alpha)$ is the correlation coefficient obtained when the south polar distribution is rotated by an angle $\alpha$ in longitude relative to the north polar distribution (Fig. 2a). Figure 2b, c shows the correlation $r(\alpha)$ and the significance of that correlation $P(\alpha)$ (see Methods) for polar hydrogen, the modelled ice stability depths and the maximum and average yearly temperatures measured by Lunar Reconnaissance Orbiter Diviner[5]. Only hydrogen shows statistically significant correlations ($>5\sigma$), with peak significance near $\alpha=180°$ ($8.3\sigma$ at 187°).

The unique, high-significance antipodal relationship suggests a fundamental connection between north and south polar hydrogen. We interpret this as evidence that the lunar spin axis has reoriented from a past spin pole position, with the expectation that volatiles will accumulate in cold traps about the instantaneous poles. Our 'palaeopole' is described by the averaged polar hydrogen maxima, corresponding to 84.5° N, 138° E in the north (84.5° S, 318° E in the south). Models of ice stability that assume an admixture of thermal environments from the current and palaeopole orientations (Fig. 1e, f) lead to a better description of the hydrogen distributions than does the current environment alone (see Methods).

This type of reorientation is known as true polar wander (TPW). In a minimum energy state, the spin axis of a planet will align with the maximum principal axis of inertia. TPW results when the maximum principal axis changes orientation owing to mass redistribution within the planet[18] (see Methods). As the principal axis changes, the planet will attempt to minimize rotational energy and reorient to align this principal axis with the spin axis. This reorientation results in motion of the spin axis with respect to the surface of the planet (although the spin axis remains fixed in inertial space). Lunar TPW of varying age, magnitude and direction has been previously suggested on the basis of topography, gravity and remnant magnetism[19–21] (see Methods); however, polar volatiles have not previously been used to infer polar wander on the Moon.

A palaeo-axis represents a previous maximum principal axis of inertia. Because the present-day lunar inertia tensor is known, we can test whether a single mass anomaly could reorient the Moon from the neutron-data-derived palaeo-axis to the present-day spin axis. Using a

[1] Planetary Science Institute, Tucson, Arizona 85719, USA. [2] Southern Methodist University, Dallas, Texas 75275, USA. [3] University of Alabama in Huntsville, Huntsville, Alabama 35899, USA. [4] Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA. [5] Earth Life Sciences Institute, Tokyo Institute of Technology, Meguro, Tokyo 152-8551, Japan. [6] University of California, Los Angeles, California 90095, USA. [7] The Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland 20723, USA. [8] Columbia University, New York, New York 10027, USA. [9] California Institute of Technology, Pasadena, California 91125, USA.

*These authors contributed equally to this work.

‡Deceased.

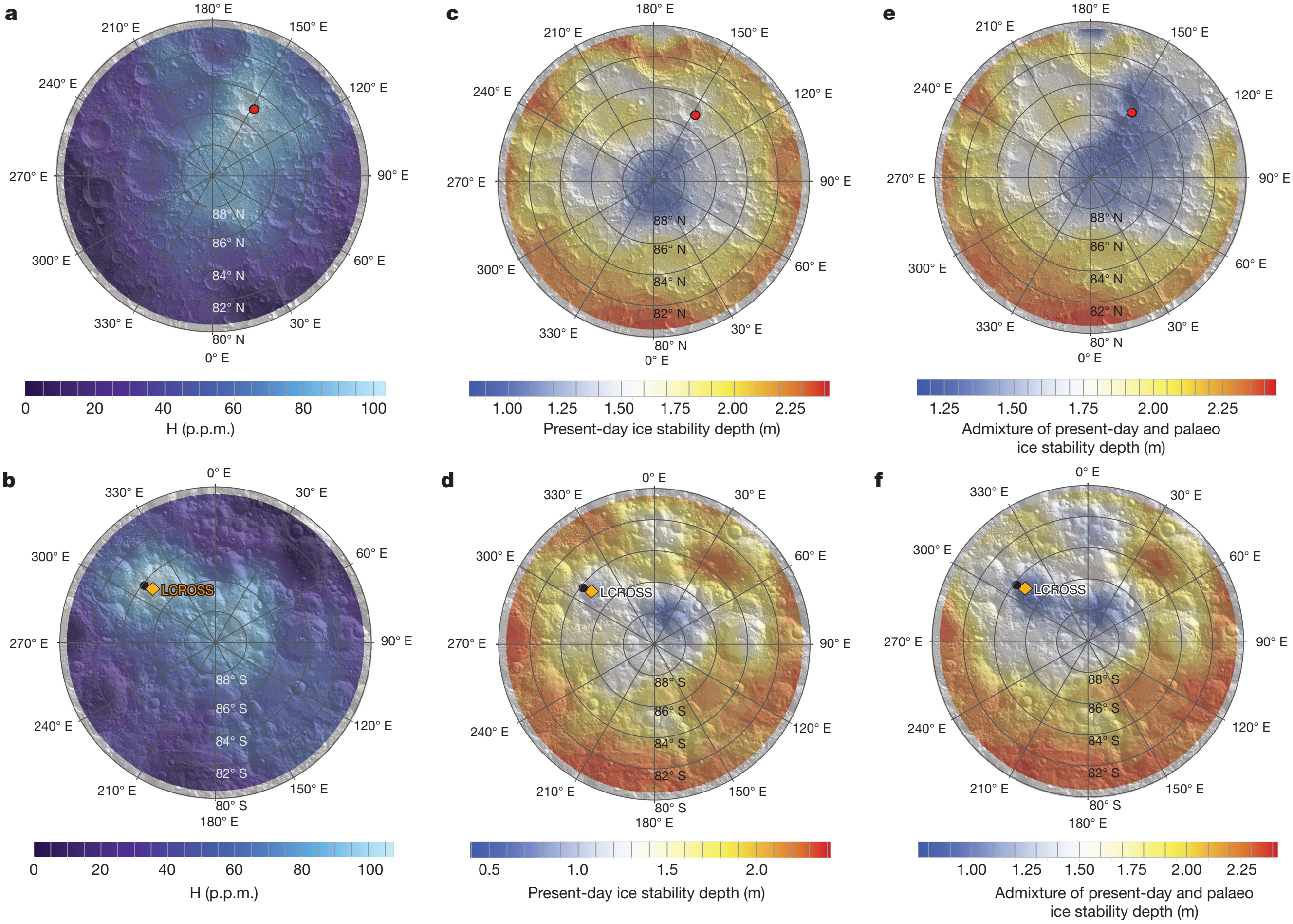


**Figure 1 | Lunar polar hydrogen and predicted ice stability fields.** **a**, **b**, Abundance enhancements of lunar polar hydrogen (H; given in p.p.m.) for the north (**a**) and south (**b**) poles[15,17]. **c**–**f**, Ice stability depths for the north (**c**, **e**) and south (**d**, **f**). Depths for the current epoch (**c**, **d**), and an admixture of current and palaeopole epochs (**e**, **f**) are shown separately. Ice stability depths are based on the model described in refs 5 and 6. In all panels, the red and black filled circles show the locations of the northern and southern hydrogen enhancement maximums, respectively, and the orange diamond shows the LCROSS impact site. All maps show quantities poleward of 80° with markers every 2°.

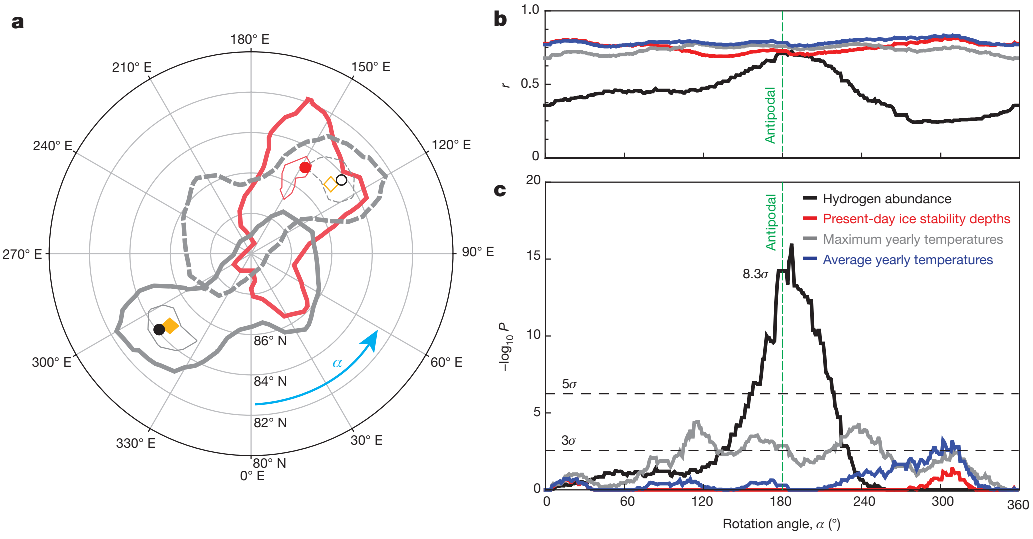


**Figure 2 | Antipodal nature of polar hydrogen.** **a**, Hydrogen abundance contours at 75 p.p.m. (thick) and 100 p.p.m. (thin) as viewed from the north pole. North (red) and south (grey; solid $\alpha = 0°$, dashed $\alpha = 180°$) spatial distributions are shown poleward of 80° with markers every 2°. The red filled circle shows the northern hydrogen enhancement maximum, the black filled circle shows the southern maximum and the filled diamond shows the LCROSS impact site; open shapes are projections onto the north. **b**, **c**, Computed correlation coefficients $r(\alpha)$ (**b**) and probability values $-\log_{10}(P)$ (**c**) versus longitudinal rotation angle $\alpha$, for hydrogen abundances (black), current ice stability depths (red), and maximum (grey) and average (blue) yearly temperatures. High values of $-\log_{10}(P)$ represent low-chance probabilities. Reference significance values at $3\sigma$ and $5\sigma$ are also shown (horizontal dashed lines). The maximum correlation is found to have $8.3\sigma$ significance.

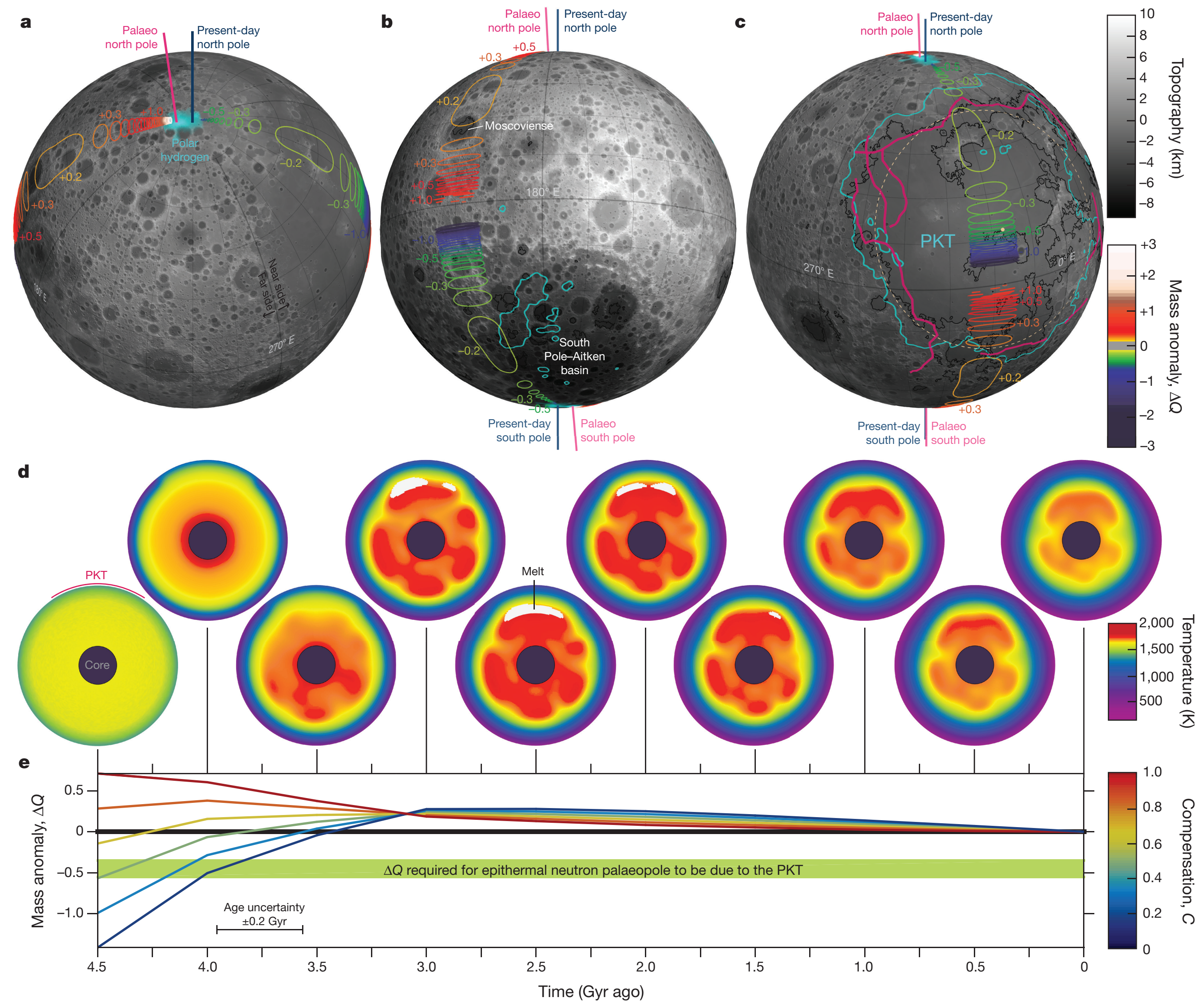


**Figure 3 | The locations of plausible mass anomalies that produce the required TPW, and the effect of the PKT. a**–**c**, Orthographic spherical projection of the lunar northern hemisphere (**a**), far side (**b**) and near side (**c**), with contours enclosing the locations where mass anomalies $\Delta Q$ of the indicated size must be centred to reorient the Moon from the epithermal neutron palaeo-axis (pink lines) to the present-day spin axis (dark blue lines). Positive mass anomalies ($\Delta Q > 0$) indicate present-day mass excesses; negative mass anomalies ($\Delta Q < 0$) indicate present-day mass deficits. Our preferred solution is a negative mass anomaly beneath the PKT (dashed tan circle and dot). For reference, LRO/LOLA (Lunar Reconnaissance Orbiter/Lunar Orbiter Laser Altimeter) topography[29] (greyscale), mare basalts boundaries from LRO/WAC (Wide Angle Camera) mapping (black lines, data available at http://wms.lroc.asu.edu/lroc/view_rdr/SHAPEFILE_LUNAR_MARE), the PKT as identified by a 3.5-p.p.m. thorium contour[30] (cyan contours) and the PKT-border gravity anomalies[11] (pink lines) are also shown. **d**, **e**, 2D cross-sections (**d**) of 3D thermochemical convection models from model W of ref. 9 that was used to determine the effective mass anomaly (**e**) associated with the PKT for different values of lithospheric compensation *C*. $C = 0$ corresponds to a rigid lithosphere; $C = 1$ corresponds to a strengthless lithosphere. This evolution is also illustrated in Supplementary Video 1.

parameter-space search, positive and negative mass anomalies of varying size were placed across the lunar surface, and changes to the lunar inertia tensor were evaluated. Reorientation in response to these mass anomalies is counteracted by the Moon's non-equilibrium degree-2 figure[21].

Figure 3a–c shows the regions in which a mass anomaly $\Delta Q = J_2^{\mathrm{MA}}/J_2$ (in which $J_2^{(\mathrm{MA})}$ is the degree-2 zonal spherical harmonic gravity coefficient of the Moon (mass anomaly); see Methods) of a given magnitude would cause a reorientation of the Moon from within 1° of the hypothesized palaeo-axis to the present-day spin axis. Both positive (mass excesses; $\Delta Q > 0$) and negative (mass deficits; $\Delta Q < 0$) anomalies are possible, although the allowed regions are limited. These regions depend only on the observed lunar inertia tensor and the location of the palaeopoles. To first order, anomalies fall on large circles connecting the present-day poles and the palaeopoles (Fig. 3a), although they deviate owing to the triaxial nature of the Moon.

Although many impact basins fall within the allowed regions, inverse modelling of the gravity fields of lunar impact basins shows that most do not have a large enough mass anomaly to produce the required reorientation[21]. Only the South Pole–Aitken basin has a large enough mass anomaly, but does not correlate with any of the allowed regions and would result in roughly orthogonal motion to our proposed wander path (Extended Data Fig. 8e).

The centre of the radiogenic-rich Procellarum KREEP Terrane (PKT; a region rich in potassium, rare-earth elements and phosphorous[7]) lies within the acceptable regions (Fig. 3c). The thermal

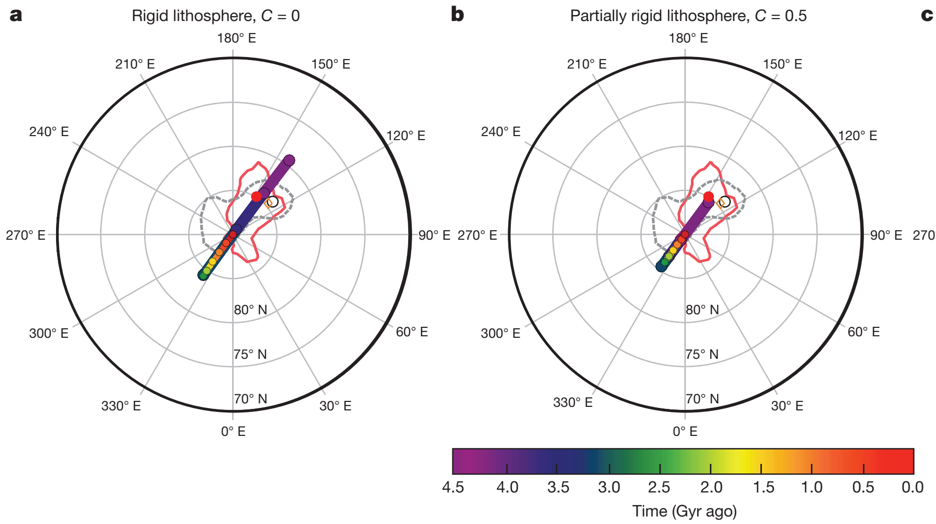


**Figure 4 | TPW due to the evolution of the PKT. a–c**, The predicted TPW path backwards through time from the present-day pole (red), owing to the thermal evolution of the PKT, for a model with KREEP material mixed within the crust (model W), assuming a rigid lithosphere ($C = 0$; **a**), a partially rigid lithosphere ($C = 0.5$; **b**) and a strengthless lithosphere ($C = 1$; **c**). In **a** and **b**, the TPW paths pass through the palaeopole, which is not forced by the model. The 1$\sigma$ uncertainty in pole position due to the rotational ambiguity of the PKT thermal anomalies are smaller than the plotted points. Other features are as in Fig. 2a.

evolution of the PKT has previously been used to explain the formation of the near-side mare basalts, the distributions of incompatible elements, the anomalous heat flow and the gravity signature of the PKT[7–11,22,23]; however, the effect of PKT thermal evolution on the moments of inertia of the Moon has not been examined. A simple model of the PKT as a spherical internal mass anomaly that spans the mantle requires only a moderate density anomaly ($|\Delta\rho| > 15\,\mathrm{kg\,m^{-3}}$; equivalent to a temperature anomaly of $|\Delta T| > 150\,\mathrm{K}$) to generate the required reorientation (see Methods). This anomaly is similar in magnitude to the thermal anomalies suggested to sit beneath the PKT[7–11,22,23].

To determine whether the thermal evolution of the PKT can produce the required reorientation, we use the 3D thermochemical convection model of ref. 9. This model (Fig. 3d) evaluates the influence of the high concentration of heat sources within the PKT on global thermal evolution. Two limiting models are considered here: heat-producing KREEP material initially beneath the crust (model 'B'), and heat-producing KREEP material initially mixed within the crust (model 'W'); see Methods. Using these models, we evaluate changes in the lunar inertia tensor and, hence, in the spin axis, as functions of time. These changes are caused by internal density variations as a result of temperature and compositional changes, and surface deformation. Figure 3e shows the mass anomaly $\Delta Q$ of our nominal PKT thermal evolution model (model W) as a function of time, for a range of lithospheric compensation states $C$. All PKT thermal evolution models produce substantial perturbations to the lunar inertia tensor, and large amounts of TPW, regardless of model parameters or compensation state (Extended Data Figs 9, 10). Models with semi-rigid lithospheres ($C < 0.5$) produce mass anomalies that are consistent with those required for PKT to be responsible for the reorientation implied by our palaeopole ($\Delta Q = -0.45$; Fig. 3c).

Figure 4 shows the chronology of three representative TPW paths derived using the nominal PKT thermal evolution for different lithospheric strengths. Starting from the present-day (non-equilibrium) lunar inertia tensor and working backwards in time allows us to track TPW due to PKT thermal evolution without making assumptions about the lunar fossil figure. TPW tracks are not forced to pass through the epithermal neutron palaeopoles; this is simply a result of placing the thermal anomaly within the PKT. As long as the lithosphere is moderately rigid, tracks will pass through the palaeopole (Fig. 4a, b, Extended Data Figs 9, 10). This result is consistent with the present geophysical state of the PKT[9] and expectations of the early rigid lunar lithosphere[24]. TPW tracks assuming a strengthless lithosphere ($C \approx 1$) do not pass through the palaeopole (Fig. 4c, Extended Data Figs 9, 10), constraining past lithospheric compensation.

Although the allowed locations and magnitudes of perturbing mass anomalies are robust (Fig. 3a–c, Extended Data Fig. 7), other geophysical processes can affect our proposed TPW, owing to the degeneracy of interpreting gravity and moments of inertia. The TPW paths and chronology depend sensitively on the compensation state of the lithosphere and on the emplacement and relaxation of near-side mare basalts. Future seismic and heat-flow measurements will constrain the current nature of the PKT thermal anomaly. Passage through the palaeopole occurs more than 3.5 Gyr ago for these representative models. Alternative chronologies that incorporate early magma-ocean evolution, mare emplacement and changes in lithospheric strength may lead to later palaeopole passage, but require additional free parameters and modelling (Extended Data Fig. 10).

Because a substantial fraction of the observed hydrogen deposits are believed to record an ancient palaeopole, they may preserve water from the early Solar System, necessitating long-term hydrogen stability. If Mercury's polar volatiles are also ancient, then lunar TPW may explain the differences between the volatile reservoirs on these two bodies (see Methods). For many TPW scenarios, the pole may not migrate substantially, water ice may always be stable in the observed locations and the detected hydrogen could simply be the remaining ice (Extended Data Fig. 3e, f), kept near to the surface by thermal migration despite impact gardening[4,13,25]. The detection of water ice near the southern palaeopole by the Lunar Crater Observation and Sensing Satellite (LCROSS) may favour this explanation[26]. If, instead, the Moon underwent a large amount of TPW, then the palaeopole may have been directly illuminated, resulting in temperatures too high for water ice. In this scenario, surficial water adsorption[27] and storage of hydrogen within grains[28] are viable mechanisms for long-term hydrogen storage (see Methods), although the spatial distribution would still point to a pre-TPW deposition in the form of ice. *In situ* measurements, sample return and high-resolution orbital geochemistry measurements could differentiate plausible TPW scenarios.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 13 July 2015; accepted 21 January 2016.**

1. Feldman, W. C. *et al.* Fluxes of fast and epithermal neutrons from Lunar Prospector: evidence for water ice at the lunar poles. *Science* **281,** 1496–1500 (1998).
2. Feldman, W. C. *et al.* Evidence for water ice near the lunar poles. *J. Geophys. Res.* **106,** 23231–23251 (2001).
3. Mitrofanov, I. G. *et al.* Hydrogen mapping of the lunar south pole using the LRO neutron detector experiment LEND. *Science* **330,** 483–486 (2010).
4. Siegler, M. A., Paige, D., Williams, J.-P. & Bills, B. Evolution of lunar polar ice stability. *Icarus* **255,** 78–87 (2015).
5. Paige, D. A. *et al.* Diviner lunar radiometer observations of cold traps in the Moon's south polar region. *Science* **330,** 479–482 (2010).
6. Paige, D. A. *et al.* Thermal stability of volatiles in the north polar region of Mercury. *Science* **339,** 300–303 (2013).
7. Jolliff, B. L., Gillis, J. J., Haskin, L. A., Korotev, R. L. & Wieczorek, M. A. Major lunar crustal terranes: surface expressions and crust-mantle origins. *J. Geophys. Res. Planets* **105,** 4197–4216 (2000).
8. Wieczorek, M. A. & Phillips, R. J. The "Procellarum KREEP Terrane": implications for mare volcanism and lunar evolution. *J. Geophys. Res. Planets* **105,** 20417–20430 (2000).
9. Laneuville, M., Wieczorek, M. A., Breuer, D. & Tosi, N. Asymmetric thermal evolution of the Moon. *J. Geophys. Res. Planets* **118,** 1435–1452 (2013).
10. Grimm, R. E. Geophysical constraints on the lunar Procellarum KREEP Terrane. *J. Geophys. Res. Planets* **118,** 768–778 (2013).
11. Andrews-Hanna, J. C. *et al.* Structure and evolution of the lunar Procellarum region as revealed by GRAIL gravity data. *Nature* **514,** 68–71 (2014).
12. Arnold, J. R. Ice in the lunar polar regions. *J. Geophys. Res. Solid Earth* **84,** 5659–5668 (1979).
13. Hurley, D. M. *et al.* Two-dimensional distribution of volatiles in the lunar regolith from space weathering simulations. *Geophys. Res. Lett.* **39,** L09203 (2012).
14. Miller, R. S., Nerurkar, G. & Lawrence, D. J. Enhanced hydrogen at the lunar poles: new insights from the detection of epithermal and fast neutron signatures. *J. Geophys. Res. Planets* **117,** E11007 (2012).
15. Miller, R.S. Statistics for orbital neutron spectroscopy of the Moon and other airless planetary bodies. *J. Geophys. Res. Planets* **117,** E00H19 (2012).
16. Miller, R. S., Lawrence, D. J. & Hurley, D. M. Identification of surface hydrogen enhancements within the Moon's Shackleton crater. *Icarus* **233,** 229–232 (2014).
17. Teodoro, L. F. A., Eke, V. R. & Elphic, R. C. Spatial distribution of lunar polar hydrogen deposits after KAGUYA (SELENE). *Geophys. Res. Lett.* **37,** L12201 (2010).
18. Matsuyama, I., Nimmo, F. & Mitrovica, J. X. Planetary reorientation. *Annu. Rev. Earth Planet. Sci.* **42,** 605–634 (2014).
19. Takahashi, F., Tsunakawa, H., Shimizu, H., Shibuya, H. & Matsushima, M. Reorientation of the early lunar pole. *Nature Geosci.* **7,** 409–412 (2014).
20. Garrick-Bethell, I., Perera, V., Nimmo, F. & Zuber, M. T. The tidal–rotational shape of the Moon and evidence for polar wander. *Nature* **512,** 181–184 (2014).
21. Keane, J. T. & Matsuyama, I. Evidence for lunar true polar wander and a past low eccentricity, synchronous lunar orbit. *Geophys. Res. Lett.* **41,** 6610–6619 (2014).
22. Zhong, S., Parmentier, E. M. & Zuber, M. T. A dynamic origin for the global asymmetry of lunar mare basalts. *Earth Planet. Sci. Lett.* **177,** 131–140 (2000).
23. Zhang, N., Parmentier, E. M. & Liang, Y. A 3D numerical study of the thermal evolution of the Moon after cumulate mantle overturn: the importance of rheology and core solidification. *J. Geophys. Res. Planets* **118,** 1789–1804 (2013).
24. Zhong, S. & Zuber, M. T. Long-wavelength topographic relaxation for self-gravitating planets and implications for the time-dependent compensation of surface topography. *J. Geophys. Res. Planets* **105,** 4153–4164 (2000).
25. Schorghofer, N. & Taylor, G. J. Subsurface migration of $H_2O$ at lunar cold traps. *J. Geophys. Res. Planets* **112** E02010 (2007).
26. Colaprete, A. *et al.* Detection of water in the LCROSS ejecta plume. *Science* **330,** 463–468 (2010).
27. Poston, M. J. *et al.* Temperature programmed desorption studies of water interactions with Apollo lunar samples 12001 and 72501. *Icarus* **255,** 24–29 (2015).
28. Starukhina, L. in *Moon: Prospective Energy and Material Resources* (ed. Badescu, V.) 57–85 (Springer, 2012).
29. Smith, D. E. *et al.* Initial observations from the Lunar Orbiter Laser Altimeter (LOLA). *Geophys. Res. Lett.* **37** L18204 (2010).
30. Lawrence, D. J. *et al.* Small area thorium features on the lunar surface. *J. Geophys. Res. Planets* **108,** 5102 (2003).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This project was supported in part (M.A.S., R.S.M. and D.J.L.) by NASA's SSERVI VORTICES node, the Lunar Reconnaissance Orbiter (M.A.S. and D.A.P.) and the NASA Lunar Advanced Science and Exploration Research (LASER) programme (J.T.K. and I.M.). We thank L. L. Hood for sharing insight on palaeomagnetic data. We dedicate this project to Arlin Crotts, who passed shortly after the completion of this paper.

**Author Contributions** M.A.S. posed the fundamental true polar wander question (with M.J.P.), identified contributors needed to address key elements of the hypothesis, and provided the near-surface thermal and ice stability models from models developed by D.A.P. R.S.M. developed and performed the analyses of neutron datasets, including the determination of hydrogen abundance and spatial distribution, as well as the multi-dataset correlation, statistical and admixture analyses. J.T.K. performed all true polar wander analyses, identified the PKT as the plausible cause of the reorientation and produced all figures. M.L. provided critical input, including results from his previously published thermal evolution model of the lunar interior. D.A.P., I.M. and D.J.L. contributed expertise related to the interpretation of thermal, gravity and orbital geochemistry science. M.J.P. and A.C. contributed expertise regarding long-term volatile stability models. M.A.S., R.S.M. and J.T.K. were the primary authors, with M.L., D.A.P., I.M., D.J.L., M.J.P. and A.C. providing additional contributions and assisting with the review of the paper.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.A.S. (msiegler@psi.edu).

## METHODS

**Determination of hydrogen abundance.** Orbital neutron spectroscopy is commonly divided into three distinct energy regimes—thermal (low energy), epithermal (intermediate energy) and fast (high energy)—each providing complimentary information about elemental abundance and distribution (spatial and depth). The process starts with fast neutrons created by cosmic-ray interactions in the lunar regolith. Elastic neutron–proton scattering causes these neutrons to rapidly lose energy, shifting some of them into the epithermal regime. Subsequent moderation and/or capture processes can further modify the flux and spectrum thereby imprinting details of the intervening material on escaping neutrons. Owing to the efficiency of the neutron energy-loss process, the epithermal regime is an especially sensitive probe of hydrogen[1,2]. Epithermal neutron deficits measured from orbit are therefore indicative of enhanced hydrogen abundances.

Proper determination of statistical significance is often exchanged for approximate methods that may be simple or reduce computation requirements. Low signal-to-noise environments require a more rigorous approach. Relevant statistical descriptions are based on particle counts, not rates, and therefore require the use of exposure distributions in addition to observed neutron count rates. Our statistical analysis approach uses a likelihood parameter $\lambda$ to characterize consistency between acquired neutron data and a hydrogen-poor (null) hypothesis on a pixel-by-pixel basis. This parameter incorporates fundamental observational details as well as the inherent uncertainties associated with counting statistics. The likelihood parameter is governed by a well-known statistical distribution ($\chi^2$) and, hence, can be used to exclude features of low or marginal significance. The null hypothesis is rejected, on a pixel-by-pixel basis, if $\lambda$ exceeds a predetermined critical value. For this work the critical value ($\lambda = 25$) was chosen to correspond to a chance probability of $5.7 \times 10^{-7}$, equivalent to a $5\sigma$ Gaussian detection with 1 degree of freedom. Additional details of the statistical analysis framework used here are found in ref. 15.

Significance maps are converted to hydrogen abundance distributions following the procedure outlined in ref. 14 (and references therein). Briefly, the statistical significance (for example, $\lambda$-statistic) is proportional to the magnitude of neutron count rate deficits, which in turn correlates directly with hydrogen abundance[1]. The relationship between neutron count rates and hydrogen abundances has been derived using Monte Carlo simulations that assume that the regolith has a composition equivalent to ferroan anorthosite[1]. Hydrogen abundance distributions for Lunar Prospector (LP), obtained following the likelihood-based analysis protocols described above, are shown in Fig. 1.

**Additional neutron data sets.** The Lunar Reconnaissance Orbiter (LRO) Lunar Exploration Neutron Detector (LRO/LEND) includes a combination of collimated and uncollimated $^3$He sensors[3,31,32] with one of the four uncollimated sensors configured for epithermal neutron detection. The collimated sensors for epithermal neutrons (CSETN) were designed to provide data with improved spatial resolution over uncollimated sensors, but low count rates and systematic background effects limit its value for confidently inferring hydrogen concentrations with high spatial resolution[14,33–36]. Because of these documented problems, the collimated LEND data are not used in this study. Uncollimated epithermal neutron data from the LEND sensors for epithermal neutrons (SETN) have been shown to have a reasonably good spatial correlation with the uncollimated LP data[14,32]. The correlations between the two uncollimated data sets, however, are not perfect, and at best only qualitative suggestions have been provided to explain discrepancies that occur in both equatorial and polar regions[32]. Similar to LP, the spatial distribution of hydrogen derived from LEND-SETN does not match the predicted locations of water ice in the present thermal environment, and shows a broad, asymmetric, slightly off-polar distribution[32]. However, there are quantitative differences between LEND and LP that are not fully understood or documented. Our confidence in the LP data is well grounded because the LP data were measured with well-characterized sensors on a boom such that backgrounds from nearby materials were both understood and minimized[37], and because the data reduction is supported by extensive documentation[38] and a well-validated comparison with modelled count rates[39]. We expect that a more detailed analysis of the LEND data could provide additional insight to the differences between these data sets. However, such an analysis is beyond the scope of this study, and we have therefore focused on the LP-derived parameters.

**Antipodal symmetry.** Two surface features are antipodal if they lie on diametrically opposite sides of a planet, such that a line connecting the two points passes through the centre of the planet. If a feature has a latitude and longitude of $(\theta, \varphi)$, then the antipode is located at $(-\theta, \varphi + 180°)$. This type of symmetry can also be referred to as an inversion, central reflection or point reflection.

In typical map projection of polar data sets (for example, Fig. 1), antipodal features do not appear simply shifted by 180°. The different handedness between the north and south polar maps results in an additional reflection. This geometry is illustrated schematically in Extended Data Fig. 1a and b, in which two antipodally symmetric features are shown in north and south polar maps, respectively. To illustrate the antipodal nature of this feature, we show both north and south features in each plot, with the antipodal feature shown as it would appear if you could view it through the Moon. In this projection, a feature rotated by an angle ($\alpha$) of 180° about the pole will exactly line up with its antipodal self (Extended Data Fig. 1c, d).

**Correlation analysis.** The Pearson product-moment correlation coefficient $r$ is used to quantify the strength of the correlations between data sets[40]. Values of this statistic occur within limits ($-1 \leq r \leq 1$) corresponding to perfect anti-correlation and correlation, respectively. A two-point correlation was implemented to operate on pixelated spatial distributions characterized by latitude and longitude. This coefficient measures similarities in relative amplitude (or shape) only, and is not used to evaluate the physics implications of the absolute neutron rates or thermal parameters.

By itself the correlation coefficient is not a good statistic for determining the quality of an observed correlation. However, the significance of differences in correlation coefficients is relevant. The Fisher $z$-transformation facilitates hypothesis testing by quantifying whether a change in some physical parameter modifies the baseline correlation between two distributions. When applied to the Pearson coefficient it stabilizes the variance[41] and can be used to determine significance. Fisher's transformation takes the form

$$z(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \tag{1}$$

and has a standard error of

$$\sigma(z) \approx \frac{1}{\sqrt{N-3}}$$

where $N$ is the number of measurements in the population. If the baseline correlation (or null hypothesis) is characterized by a coefficient $r_0$ and a second correlation by $r$, then the two-sided significance of the difference between the two measured coefficients ($\Delta z = |z(r_0) - z(r)|$) is

$$P = \operatorname{erfc}\left(\frac{\Delta z\sqrt{N-3}}{\sqrt{2}}\right) \tag{2}$$

where erfc($x$) is the complementary error function[41]. The Fisher transformation also enables determination of confidence levels. Because $z$ can be approximated by a normal distribution with known variance, a 90% confidence interval is given by

$$z - 1.64\sigma(z) \leq z \leq z + 1.64\sigma(z)$$

and application of the inverse Fisher transform yields the relevant confidence intervals on the measured correlation coefficient $r$.

The determination of significance assumes that polar map pixels are independent. A globally mapped, equal-area pixel size was selected to match the ~45-km spatial resolution of acquired neutron data[14], but total independence cannot be assured. This results in each polar region containing 364 equal-area pixels. Of those, only 248 (123 in the north and 125 in the south) meet the statistical threshold requirement discussed below ($\lambda = 25$). The number of pixels ($N$) used to evaluate significance was independent for each $\alpha$ and includes only those pixels common to the unrotated case ($\alpha = 0°$); at peak significance $N = 236$.

Using equation (1), the peak in Fig. 2b corresponds to $\Delta z = 0.548$ (with $r = 0.728$ and $r_0 = 0.356$). Substituting these values into equation (2), we obtain $-\log_{10}(P) = 16.2$, which corresponds to the peak in Fig. 2c and is equivalent to about $8.3\sigma$. Even if the spatial distributions are oversampled by a factor of two—an extreme exaggeration that reduces the number of pixels to $N = 118$—the observed antipodal correlation is still significant, with a chance probability $<10^{-9}$, which exceeds a $5\sigma$ threshold.

Random processes (noise) will degrade any observed correlations. Therefore, investigating the dependence of inter-polar correlations on the likelihood parameter $\lambda$ is instructive because it serves as a proxy for statistical significance of features. North and south polar hydrogen distributions show evidence for a strong near-antipodal relationship. Extended Data Fig. 2 shows peak significance $-\log_{10}(P)$ as a function of $\lambda$. A reduction in correlation significance as low-$\lambda$ (low statistical significance) features are included is evident. Such a trend is expected if the features identified above the critical threshold ($\lambda \geq 25$) are real, and those below are dominated by statistical fluctuations.

**Surface thermal model to examine ice stability.** The central argument for the existence of lunar volatiles is based on thermal modelling[42]. The forward thermal model presented here (for example, Figs 1c, d, 3 and 4) is an updated version of that presented in refs 5 and 6. The thermal model we use is intended to be the simplest model that can reproduce the major features of the LRO Diviner south polar observations. Updates to the model of ref. 5 include the use of polar meshes and an updated model of the Sun and its ephemeris, as detailed below.

Polar meshes were modified to reproduce thermal conditions under the assumption of a past spin by transforming the polar stereographic '*z*' coordinate to appropriately represent the change in polar position. Our offset figures were created with a map shifted to have the poles at 84.5° N, 138.6° E and 84.5° S, 318.6° E, which correspond to simple averages of the two polar hydrogen maxima.

We model the Sun using a triangular mesh consisting of 128 triangles whose radiance decreases with distance from the centre of the Sun according to the solar limb darkening curve[5,6]. The location and distance of the Sun relative to the Moon as a function of time is determined using the DE421 JPL Planetary Ephemeris.

The full-resolution thermal-model results for ice stability depths in Fig. 1 are presented in Extended Data Fig. 3a–d. A version of Extended Data Fig. 3b has previously been published in ref. 5, but the remaining models are new. These models show where ice would be stable from sublimation at a rate of 1 mm Gyr$^{-1}$ assuming a regolith cover. These depths have been shown to be consistent with radar[43,44] and neutron-spectrometer-derived depths on Mercury[45].

Volatiles should collect in the most thermally stable environments—permanently shadowed regions. Using a thermal and ice stability model[4–6], the location of possible volatile reservoirs can be identified for different polar axis locations.

The model outputs for the current lunar orientation are shown in Fig. 1. To facilitate direct comparisons with the hydrogen distributions the model outputs have been degraded to a spatial resolution of 30 km from the original 0.5 km to approximate the spatial resolution of the LP Neutron Spectrometer instrument in its low-altitude orbit. Water-ice stability depths for the current orientation, the proposed palaeopole orientation and an admixture between the two (at about 30-km resolution) are shown in Extended Data Fig. 4, which repeats parts of Fig. 1 for clarity.

Admixtures of the present-day- and palaeo-axis model results (Fig. 1e, f; Extended Data Fig. 4c, f) are better correlated to the neutron data than is the present-day model alone. A given admixture is a reasonable descriptor if the corresponding correlation between it and the hydrogen distribution improves; here, statistical significance is measured relative to the correlation with a pure current spin-axis thermal model.

For the north polar region, a present-day-only model (Fig. 1c, Extended Data Fig. 4a) is excluded at the 90% confidence level and the best descriptor is a 57%–43% admixture of current- and palaeo-axis hypotheses, respectively. The south polar region (Fig. 1d, Extended Data Fig. 4d) is consistent with a pure current spin-axis hypothesis, although the optimum north pole admixture (the 57%–43% mixture) is allowed at the 90% confidence level. This strengthens the argument that the identified longitudinal bias is related to topographic and thermal effects on hydrogen.

We caution not to over-interpret the thermal analysis because it is an approximation that incorporates only two unique polar-axis locations. A more rigorous analysis must fully account for the TPW path and chronology. However, if temperature is the fundamental parameter driving volatile retention, then this approximation provides useful insights and additional support for our hypothesis of a palaeo-axis and TPW migration. Given higher-resolution neutron measurements[46] and advances in polar crater chronology[47], it may be possible to use comparisons between neutron data and crater age to help constrain the timescale of the suggested lunar TPW. If certain craters did not exist at the time of hydrogen deposition, then they will plausibly remove near-surface hydrogen-rich materials, setting a lower limit on hydrogen age. Conversely, if hydrogen is found to be associated with relatively young craters (about 2–3.5 Gyr), then it will set an upper limit on the age of hydrogen emplacement and constrain many TPW models. Evolution of lunar obliquity can also influence volatile survivability and its spatial distribution and may inform this timeline[4,12,48,49].

**True polar wander.** Changes in the spin axis of a planetary body fall into two categories[50–52]: changes in obliquity and true polar wander (TPW) (Extended Data Fig. 5).

The first category involves changes in the orientation of the spin axis in inertial space (that is, changing the position of the spin axis with respect to the celestial sphere), but not with respect to the surface of the planet (Extended Data Fig. 5b). In other words: the obliquity of the planet (the angle between the planet's spin axis and the planet's orbit normal) changes. Changes of this type result from external torques acting on the planet that can alter the planet's angular momentum (both in magnitude and direction). For planets, the most notable torques are tidal torques from satellites, the Sun and other planets. Precession and nutation are well-known examples of this form of spin evolution for the Earth, as are Cassini state transitions for the Moon and Mercury[53,54]. Spin evolution of this type can have large influences on the stability of ice at the lunar poles[4,12,48,49]. In general, near-zero obliquity is required for ice stability at the poles.

The second category of changes in planetary spin axes are those that change the orientation of the spin axis with respect to the surface of the planet, but do not change the position of the spin axis in inertial space (Extended Data Fig. 5c). This reorientation of the planet with respect to the spin axis is generally referred to as TPW[18,55,56]. Changes of this type are due to changes in the mass distribution within a planet or its hydrosphere/atmosphere. Redistribution of mass within the planet alters its inertia tensor. In a minimum energy rotation state, the rotation axis will be aligned with the maximum principal axis of inertia. If the mass redistribution changes the direction of the maximum principal axis, then the planet will reorient to keep the maximum principal axis aligned with the spin axis. Thus, to an outside (inertial) observer, the surface of the body appears to reorient with respect to the spin axis and maximum principal axis of inertia—as long as the changes in the mass distribution occur slowly with respect to the free precession period of the planet. If the changes in the inertia tensor are rapid (as might happen in the aftermath of a giant asteroid impact), the planet will enter an ephemeral period of non-principal axis rotation until the planet dissipates enough energy to return to principal axis rotation[58]. For most non-catastrophic geologic processes (for example, mantle convection and isostatic relaxation of topography), it is generally safe to assume that the planet always remains in principal axis rotation. TPW has been directly measured for the Earth, in the form of periodic TPW (driven by seasonal variations in atmospheric pressure, oceanic currents and ice loading) and secular TPW (driven by post-glacial rebound and mantle convection)[50]. Beyond Earth, TPW has been inferred for a variety of planetary bodies, including the Moon[19–21,59–64], Enceladus[65,66], Europa[66–68] and Mars[69,70] (see ref. 18 for a review). Because TPW does not change the orientation of the planet's spin vector in inertial space, the instantaneous spin pole can remain a volatile cold trap (for sufficiently small obliquities).

**The many palaeopoles of the Moon.** Our epithermal neutron palaeopole is not the first palaeopole proposed for the Moon. Extended Data Fig. 6 summarizes all previously proposed lunar palaeopoles. Lunar palaeopoles can be subdivided into three distinct categories on the basis of the data set used to identify them: (1) palaeomagnetic poles, (2) fossil-figure poles determined from long-wavelength topography or gravity, and (3) palaeopoles inferred from the distribution of polar volatiles (proposed for the first time here). Here we summarize these methods and the associated difficulties.

The first lunar palaeopoles were inferred from orbital surveys of crustal magnetic anomalies from Apollo 15 and 16 sub-satellites[59,60], and have subsequently been measured to higher precession with Lunar Prospector[61] and Kaguya[19,63,64] observations. These magnetic anomalies can be fitted with source models of varying prescription, and a local dipole magnitude and orientation can be determined. Assuming that this local dipole is a frozen remnant from a global, body-centred core dynamo field, the geometry of this local field can be used to infer a palaeomagnetic pole (that is, the surface location where the magnetic dipole intersects the surface). Under the assumption that the dipole is aligned with spin axis of the Moon, this palaeomagnetic pole is then a record of the spin pole at the time at which the magnetic anomaly formed.

There are several difficulties with interpreting palaeomagnetic poles. First, not all magnetic anomalies trace global dynamos. Large-scale impacts generate transient magnetic fields that can be different from any core dynamo existing at that time. Many deposits associated with magnetic anomalies (particularly those associated with impact ejecta, or features antipodal to large basins) may have experienced rapid shock-remnant magnetization during these transient fields, and thus may not accurately trace a core dynamo. To determine a true magnetic palaeopole, it is necessary to identify deposits that cooled slowly, well after the dissipation of any transient field (that is, thermoremnant magnetization). Identifying these deposits is difficult, and has been done convincingly only for a few magnetic anomalies[61,64]. Although disentangling shock-remnant and thermoremnant anomalies is difficult, it is still curious that many magnetic anomalies cluster into two groups: one near the present-day spin-pole, and one in the far-side mid-latitudes[19]. The second major difficulty with interpreting palaeomagnetic poles is that they may not accurately trace the spin axis of a planet. This is the case on Earth, where the magnetic pole is misaligned with the spin pole by about 10°. Future work will need to investigate the formation and evolution of the lunar dynamo, in 3D, to determine how large of a misalignment is possible.

There have been some attempts to infer palaeomagnetic poles from analysis of remnant magnetism in samples returned from the Apollo missions. Because the original orientation of these samples is unknown, it is not possible to completely describe the field geometry at the time these samples acquired their magnetizations—however, it is possible to infer the palaeolatitude of the samples on the basis of the orientation of the remnant field with respect to the sample's magnetic fabric, which is used as a proxy for palaeohorizontal. Analysis of multiple samples from multiple Apollo landing sites has been used to infer palaeomagnetic poles[62].

The second types of palaeopoles are those inferred from measurements of the Moon's long-wavelength topographic shape and gravitational field—the so-called 'fossil figure'. Following the Moon's formation and differentiation, the Moon was largely molten, and probably possessed a triaxial figure in equilibrium with the tidal and rotational potential of its early orbit. Eventually, the Moon cooled, and developed an elastic lithosphere capable of supporting this primordial, fossil triaxial figure over geologic time. The axis associated with the maximum principal moment of inertia of this figure would represent the palaeopole at the time that the elastic lithosphere formed. This fossil figure was preserved even as the Moon migrated to larger radial distances from Earth and the tidal and rotational potentials decreased.

Although it is possible to directly measure the Moon's present-day figure and its associated pole (quantified by degree-2 gravity and topography, and libration measurements), it is non-trivial to measure the primordial figure. Giant impact basins (particularly the South Pole–Aitken basin) and other large-scale geologic processes alter the Moon's figure and obscure the true fossil figure. Garrick-Bethell *et al.*[20] and Keane and Matsuyama[21] have developed two different methods for isolating this fossil figure. A critical comparison of these two works is beyond the scope of this paper, but both suggest that the fossil figure has reoriented by 15°–30° (although in different directions).

Although there is substantial scatter in the lunar palaeopoles reported in the literature (Extended Data Fig. 6), future work might be able to synthesize these data sets into a cohesive history of lunar TPW. Studies of the lunar fossil figure[20,21] should provide the 'initial' spin pole of the Moon. Paleomagnetic poles probably trace the lunar pole during the subsequent 1 Gyr, when the core dynamo was active[71]. Because polar volatiles are stable only during near-zero (roughly $<12°$) obliquity, polar volatiles probably trace polar wander only after the highly uncertain Cassini-state transition[49]. Although polar volatiles may not be able to trace the earliest episodes of lunar TPW, they have the distinct advantage of being capable of tracing small amounts of polar wander, late in lunar history.

**Identifying the mass anomaly responsible for the epithermal neutron palaeopole.** Under the assumption that the epithermal neutron palaeopole (north pole: 84.9° N, 147.9° E; south pole: 84.1° S, 309.4° E) is a former rotational palaeopole, and thus a former maximum principal axis of inertia, we ask the question: what mass anomaly would be required to reorient the Moon from this palaeopole to its present-day spin pole (0° N/S)? Phrased in terms of inertia tensors, this question is equivalent to $I = I_{\text{palaeo}} + I_{\text{MA}}$, in which $I$ is the present-day lunar inertia tensor, $I_{\text{MA}}$ is the inertia tensor of some arbitrary mass anomaly and $I_{\text{palaeo}}$ is an undetermined inertia tensor with the maximum principal axis of inertia aligned with the epithermal neutron palaeopole. The goal here is to find all possible $I_{\text{MA}}$ that satisfy this condition.

The present-day lunar inertia tensor $I$ can be determined directly from a combination of degree-2 spherical harmonic gravity coefficients, $J_2$ ($-C_{20}$) and $C_{22}$, and libration parameters, $\beta$ and $\gamma$, and is well constrained[21,57,72,73]. In a principal-axis reference frame, the lunar inertia tensor can be written

$$I = \begin{bmatrix} A & 0 & 0 \\ 0 & B & 0 \\ 0 & 0 & C \end{bmatrix}$$

in which $A$, $B$ and $C$ are the minimum, intermediate and maximum principal moments of inertia. Following ref. 57, it is convenient to define these principal moments in terms of their departures from the mean moment of inertia

$$\frac{I}{MR^2} = \frac{A+B+C}{3MR^2} = \frac{2}{3}\left(-J_2 + \frac{6C_{22}}{\gamma}\right) = 0.39298$$

in which $M$ and $R$ are the mass and radius of the Moon, respectively. $A$, $B$ and $C$ can then be written

$$\frac{C-I}{MR^2} = \frac{2}{3}J_2 = 135.619 \times 10^{-6}$$
$$\frac{B-I}{MR^2} = -\frac{1}{3}J_2 + C_{22} = -23.019 \times 10^{-6}$$
$$\frac{A-I}{MR^2} = -\frac{1}{3}J_2 - 2C_{22} = 135.619 \times 10^{-6}$$

The inertia tensor of an arbitrary mass anomaly $I_{\text{MA}}$ depends strongly on the assumed location, geometry and mass distribution of the perturbing mass anomaly. However, if we consider the limiting case of an axisymmetric anomaly centred on the north pole of the planet (such that the symmetry axis of the anomaly, and the $z$ axis are aligned), then the inertia tensor of the anomaly is reduced to a single parameter. This is because an axisymmetric anomaly centred on the north pole will contribute only to $J_2$, no other degree-2 spherical harmonic gravity coefficients, owing to symmetry. Following ref. 50, we then relate the degree-2 gravity of the mass anomaly located on the pole directly to an inertia tensor

$$I^{*}_{\text{MA}} = MR^2 \begin{bmatrix} \frac{I}{MR^2} - \frac{1}{3}J_2^{\text{MA}} & 0 & 0 \\ 0 & \frac{I}{MR^2} - \frac{1}{3}J_2^{\text{MA}} & 0 \\ 0 & 0 & \frac{I}{MR^2} + \frac{2}{3}J_2^{\text{MA}} \end{bmatrix}$$

in which $J_2^{\text{MA}}$ is the degree-2 zonal spherical harmonic coefficient associated with the mass anomaly when centred on the north pole (aligned with the positive $z$ axis). Because we are concerned only with the orientation of the principal axes of inertia (the maximum of which is presumed to be associated with a palaeopole), the mean moment of inertia can be neglected. The mean moment of inertia is spherically symmetric and does not control the orientation of the principal axis of inertia. (Stated another way: the mean moment of inertia affects the eigenvalues of the inertia tensor, but not the eigenvectors.) $I^{*}_{\text{MA}}$ is the inertia tensor for the case in which the mass anomaly is located on the north pole (with the symmetry axis aligned with the positive $z$ axis). To determine the inertia tensor for a mass anomaly located anywhere on the Moon, we rotate the inertia tensor: $I_{\text{MA}} = \mathcal{R} I^{*}_{\text{MA}} \mathcal{R}^{\top}$, in which $\mathcal{R}$ is a rotation matrix to rotate the mass anomaly from the north pole to an arbitrary latitude and longitude and $\mathcal{R}^{\top}$ is the transpose of $\mathcal{R}$. Ultimately, $I_{\text{MA}}$ is simply a function of $J_2^{\text{MA}}$ and the position (latitude and longitude) of the anomaly. For simplicity, we define the quantity $Q = -J_2^{\text{MA}}/J_2$, with $J_2$ the degree-2 zonal gravity harmonic measured by GRAIL[74]: $J_2 = -C_{20} = 203.2133 \times 10^{-6}$, in unnormalized spherical harmonic coefficients. The negative sign forces $Q$ to be positive for positive mass anomalies and negative for negative mass anomalies.

To determine the possible locations and magnitudes of perturbing mass anomalies that could be responsible for the observed epithermal neutron palaeopole, we performed a parameter-space survey investigating the effect of placing mass anomalies of various sizes ($Q$) across the surface of the Moon. For each test case, we determined the palaeo inertia tensor: $I - I_{\text{MA}} = I_{\text{palaeo}}$.

We determined the orientations of the principal axes of inertia by evaluating the eigenvalues and eigenvectors of $I_{\text{palaeo}}$. We then measured the mean angular separation between the maximum principal axis of inertia and the epithermal neutron north and south poles. Extended Data Figure 7a–d shows example slices of this parameter-space search for positive and negative mass anomalies. The regions that can drive the required reorientation to within the measured uncertainty (approximately 1°) are limited. Figure 3a–c and Extended Data Fig. 7e show the acceptable regions in which a mass anomaly of a range of sizes ($Q$) could produce the required reorientation to within 1°.

**Simple physical models for producing the required reorientation.** Lunar impact basins, uncompensated topography and mare basalts can have a substantial contribution to the inertia tensor of the Moon[21,58,75]. To determine if these features were possibly responsible for the reorientation that is required to explain the epithermal neutron palaeopoles, we considered a simple case of a spherical cap of uniform surface density (Extended Data Fig. 8a). Extended Data Figure 8b shows $Q$ for spherical caps as a function of surface density (which, assuming a material density, can be converted into an equivalent, uncompensated material thickness) and cap radius. For the typical sizes of large impact basins (radii of $<15°$), required mass anomalies (Fig. 3a–c; $|Q| > 0.2$) would be equivalent to $>5$ km of uncompensated topography (either a topographic excess or depression, depending on the sign of the surface density). This magnitude of uncompensated topography or mare basalts is not observed in any lunar impact basin. In the following section, we exclude impact basins and mare basalts in a more rigorous manner.

Internal mass anomalies, including mantle plumes or lateral variations in composition or density, can also have a contribution to the Moon's inertia tensor. For simplicity, we considered a simple spherical mass anomaly, spanning from the outer core to the lunar crust, with an arbitrary density contrast (Extended Data Fig. 8c). In this case, $Q$ for this simple internal anomaly is dependent only on the density contrast, as shown in Extended Data Fig. 8d. The smallest required mass anomalies (Fig. 3a–c; $|Q| \approx 0.2$) would be equivalent to density anomalies of only $|\Delta\rho| \approx 10\,\text{kg}\,\text{m}^{-3}$. If these density anomalies are thought to arise from temperature variations, this would be equivalent to $|\Delta T| \approx 100$ K (assuming a 3,300 $\text{kg}\,\text{m}^{-3}$ mantle density and a volumetric coefficient of thermal expansion of $3 \times 10^{-5}$)[76]. Temperature anomalies of this magnitude are easily generated in thermal evolution models of the PKT[8–11]. This drives our detailed investigation of the TPW potential of the PKT.

**The contribution of impact basins to the lunar inertia tensor.** To determine whether impact-basin features can produce the mass anomalies required to explain the epithermal neutron palaeopole, we used the method of ref. 21 to isolate the degree-2 gravity field of these features. Extended Data Figure 8f shows the best-fit mass anomaly ($Q$) associated with each of the 32 largest lunar impact basins.

From Extended Data Fig. 8f it is clear that most lunar impact basins have a small contribution to the degree-2 gravity field of the Moon—with the exception of the South Pole–Aitken basin, and its associated ejecta blanket. All other impact basins have $|Q| < 0.2$, which is the smallest possible value of $Q$ that can reorient the Moon enough to explain the epithermal neutron palaeopoles (Fig. 3a–c). The only large impact basin that is located in a place that could potentially reorient the Moon in the necessary direction is Moscoviense (27° N, 148° E). For it to cause the observed reorientation, Moscoviense would need to be a present-day positive mass anomaly, with $Q \approx +0.22$ (Fig. 3a–c). From the inverse modelling of this basin's gravity field, we find that Moscoviense is a net negative mass anomaly, with $Q < 0.1$. Thus, even the favourably located Moscoviense impact basin is not capable of causing the required reorientation.

Lunar impact basins tend to have a negligible contribution to degree-2, owing to the detailed structure of their gravity fields. Large lunar impact basins frequently possess large, central, positive free-air anomalies (so-called 'mascons'[77]), surrounded by a broad, negative free-air anomaly collar resulting from the deposition of ejecta and thickening of the crust[78]. This alternating positive/negative 'bull's-eye' pattern results in an almost net-zero contribution to the degree-2 gravity field[21]. It is possible that impact basins had more substantial contributions to degree-2 shortly after they formed, and before the formation of the central mascon, due to viscoelastic relaxation, mantle-flow, and cooling and contraction of the impact melt pool; however, this would be a transient stage lasting less than 30 Myr (ref. 78). It is unlikely that all of the observed hydrogen deposits formed in such a short time-span. Furthermore, if large impact basins were responsible, then we would expect several sets of antipodal epithermal neutron deposits, rather than just one.

Although the South Pole–Aitken basin and its associated global ejecta blanket easily produce mass anomalies comparable to those required to explain the epithermal neutron palaeopoles[21] (Extended Data Fig. 8f), it is not at the proper location to reorient the Moon in the necessary direction (Fig. 3b). In fact, the location of the South Pole–Aitken basin is incompatible with the observed epithermal neutron palaeopole. Extended Data Figure 8e illustrates the range of possible palaeopoles for both the South Pole–Aitken basin and the PKT, for a wide range of mass anomalies centred on each feature (the entire parameter space of Extended Data Fig. 8b). The latitude and longitude of the perturbing mass anomaly immediately restricts the possible locations for a palaeopole. The set of possible palaeopoles for PKT naturally passes through the epithermal neutron palaeopoles, whereas the possible palaeopoles associated with the South Pole–Aitken basin are nearly orthogonal to the observed reorientation. Thus, the South Pole–Aitken basin cannot be responsible for the observed epithermal neutron palaeopoles (although asymmetries in the impact basin and associated ejecta blanket[79] may complicate this picture).

**Evolution of the lunar inertia tensor due to the formation and evolution of the PKT.** To determine the reorientation of the Moon due to the thermal evolution of the Procellarum KREEP Terrain (PKT), we used the 3D thermochemical convection models of ref. 9 (see ref. 9 for the details of the model). Here, we focus on how we use these models to determine the TPW history of the Moon.

The PKT thermal models consist of a 3D spherical grid, with 20-km radial resolution and 60-km lateral resolution. The radial grid runs from the core–mantle boundary (at a radius of $R_C = 390$ km) to the Moon's surface ($R_P = 1{,}740$ km). At each volume element within the model domain, the density varies owing to thermal expansion/contraction; in the bulk composition, it varies owing to partial melting and subsequent melt migration. We determine the full inertia tensor of the model by summing the contribution of each volume element

$$I_{xx} = \sum_i \rho_i V_i (y_i^2 + z_i^2)$$
$$I_{xy} = -\sum_i \rho_i V_i x_i y$$

and similarly for the other components of the inertia tensor ($I_{yy}$, $I_{zz}$, $I_{xz}$, $I_{yz}$). Here, $V_i$ is the volume of the $i$th grid element and $\rho_i$ is the density, which varies with time. In these calculations, we take PKT to be located along the positive $z$ axis.

For TPW, it is not only important to consider density variations within the body, but also surface deformation in response to the temperature evolution at depth. As the mantle heats up, the surface will be uplifted in response to the thermal expansion of the mantle. Depending on the magnitude of this surface compensation, it is possible for PKT to act as either a net negative anomaly ($Q < 0$; if the thermal anomaly at depth dominates) or a net positive anomaly ($Q > 0$; if the topographic uplift dominates). Our PKT models do not directly take into account changes in surface topography due to thermal evolution at depth. To address this, we followed the approach used in ref. 9 and calculated the amount of surface uplift *a posteriori* by determining the amount of topography necessary to balance the thermal expansion/contraction of the mantle at depth. For each radial column within the model domain, we determined the initial integrated mass within that column. As the interior warms owing to the evolution of PKT, this results in an overall decrease in density in the column, which, in an incompressible model without surface flexure, leads to a small decrease in the integrated mass within the column. If we assume that the lithosphere can perfectly compensate for this change in density (which would occur only if the lithosphere was completely strengthless), then we add this missing mass back into the model at the uppermost radial volume element within the column. This added mass is a proxy for the topographic uplift resulting from this interior changes in density. Because real planetary lithospheres are not strengthless, and instead possess some rigidity, we modulated this correction by a factor we term the 'compensation state' $C$. If $C = 1$, then we add in the complete mass correction corresponding to a strengthless lithosphere. If $C = 0$, then we do not add in any mass correction, which would correspond to a completely rigid lithosphere, incapable of deforming in response to the interior thermal expansion. Thus, the total inertia tensor $I_{\mathrm{PKT}}$ from the thermal model is $I_{\mathrm{PKT}} = I_{\mathrm{interior}} + CI_{\mathrm{topography}}$, in which $I_{\mathrm{interior}}$ is the inertia tensor that results from summing up the contribution of each volume element within the model and $I_{\mathrm{topography}}$ is the inertia tensor that results from the mass due to this dynamic topography in the upper-most grid cell. For all cases, we normalize the final total inertia tensor to the observed mass and radius of the Moon.

From the inertia tensor, it is possible to directly calculate spherical harmonic gravity coefficients[50]

$$C_{20} = -J_2 = -\frac{I_{zz} - \frac{1}{2}(I_{xx} + I_{yy})}{MR^2}$$
$$C_{21} = -\frac{I_{xz}}{MR^2}$$
$$C_{22} = -\frac{\frac{1}{4}(I_{xx} - I_{yy})}{MR^2}$$
$$S_{21} = -\frac{I_{yz}}{MR^2}$$
$$S_{22} = -\frac{\frac{1}{2}I_{xy}}{MR^2}$$

For the case with the PKT centred on the $z$ axis, the degree-2 gravity field associated with PKT is described primarily by $C_{20}$, owing to symmetry. Although there is some power in the other spherical harmonic gravity coefficients, $C_{20}$ is the most important. The inertia tensor is uniquely related to degree-2 gravity coefficients (and only degree-2 gravity coefficients).

In our parameter-space search for possible perturbing mass anomalies (Fig. 3a–c), we assume that the Moon used to have its spin axis at a different location (possibly at the epithermal neutron palaeopole) and was subsequently reoriented to the present-day spin pole. Phrased differently, we assume that the perturbing mass anomaly is still present, and still contributes to the observed lunar inertia tensor and degree-2 spherical harmonic gravity coefficients. Thus, to determine the relative importance of the PKT, it is more useful to define the change in the mass-anomaly size with respect to its present value: $\Delta Q = Q(t) - \Delta Q(0\,\mathrm{Gyr\ ago})$. This $\Delta Q$ is the relevant quantity for the parameter-space survey in Fig. 3a–c, and determines how much the Moon could have reoriented in the past, with respect to its present-day orientation. A positive $\Delta Q$ indicates the presence of a positive mass anomaly (mass excess) with respect to the present state; a negative $\Delta Q$ indicates the presence of a negative mass anomaly (mass deficit) with respect to the present state. $Q$ and $\Delta Q$ for two end-member PKT thermal anomalies are shown in Extended Data Figs 9, 10. The nomenclature 'W' and 'B' are shortened from '0LW' and '0LB' adopted from ref. 9, in which '0' denotes low radiogenic mantle composition and 'L' denotes the larger (in diameter) of two test cases, 'W' denotes KREEP within the crust and 'B' denotes KREEP below the crust.

**True polar wander due to the PKT.** To determine how the Moon would reorient under the thermal evolution of the PKT, it is necessary to first reorient the PKT inertia tensor so that it is properly aligned with the approximate centre of the PKT (18° N, 334° E). This can be done by either directly rotating the inertia tensor or rotating the spherical harmonic gravity coefficients via the spherical harmonic addition theorem[57,80].

To determine the TPW path predicted from the thermal evolution of the PKT, it is necessary to determine the location of the maximum axis of inertia as a function of time. The location of this maximum axis of inertia is defined as the palaeopole. To calculate this TPW path for any PKT thermal model, we first assume that the present-day, observed lunar inertia tensor is the sum of the inertia tensor from the final time-step of the thermal evolution models ($t = 0$ Gyr ago), and some non-hydrostatic component (including other impact basins, mascons, the fossil figure and so on) $I_{NH}$: $I = I_{PKT}(0\,\mathrm{Gyr\ ago}) + I_{NH}$. In this calculation, we remove the hydrostatic component of the present-day, observed lunar inertia tensor[73]. Although much of the geologic history outside of the PKT is buried within $I_{NH}$, it is important to note that most of the other geologic processes on the Moon (for example, impact basins and mare basalts) have negligible contributions to the lunar inertia tensor[21] (Extended Data Fig. 8f). The most substantial other contributors are the South Pole–Aitken basin and its global ejecta blanket[21], and the Moon's fossil figure—the remnant rotational and tidal bulge, preserved from when the Moon's lithosphere cooled sufficiently to support long-term deformation. Although the nature of the fossil figure is debated[20,21], and the formation of the South Pole–Aitken basin is still poorly understood[79], both of these events would have occurred very early in lunar history, probably predating the initial conditions of our PKT thermal model. Thus, we do not expect $I_{NH}$ to change appreciably during the course of lunar history, but rather expect only negligible perturbations due to the formation of impact basins with time. Because $I$ is known and $I_{PKT}(0\,\mathrm{Gyr\ ago})$ is inferred from our PKT thermal evolution models, we rearrange the above equation ($I = I_{PKT}(0\,\mathrm{Gyr\ ago}) + I_{NH}$) to determine $I_{NH}$.

By isolating the non-PKT, non-hydrostatic component of the lunar inertia tensor, we then determine the inertia tensor as a function of time from our PKT thermal models: $I(t) = I_{PKT}(t) + I_{NH}$. The palaeopole can be calculated at any time-step in the model by taking the inertia tensor at that time, evaluating the eigenvalue problem and identifying the orientation of the maximum axis of inertia. Figure 4 and Extended Data Figs 9 and 10 show representative TPW tracks calculated using this method. Supplementary Video 1 shows an example of this TPW for our nominal model (model W; $C = 0$), as viewed from an outside observer. As a consequence of our definition of $I_{NH}$, the TPW track will always end at the present-day rotation pole. However, the TPW track is not forced to pass through the epithermal neutron palaeopole, although this happens frequently, owing to the placement of the PKT model at the PKT.

The thermal evolution of the PKT is not completely axisymmetric about the centre of the PKT, owing to the 3D nature of the problem. This results in intermediate and minimum principal axes of inertia that are not quite equal ($I_{xx} \neq I_{yy}$), in addition to small, non-zero off-diagonal terms in the inertia tensor ($I_{xy} \neq I_{xz} \neq I_{yz} \neq 0$). These terms can have a small effect on the orientation of the maximum principal axis of inertia derived using the above method. To account for this variation, we rotate the PKT anomaly about the vector aligned with the PKT, and repeat the analysis for all possible PKT rotation angles. Error bars in our TPW paths (for example, in Fig. 4) indicate the $1\sigma$ uncertainty in the palaeopole position that results from this effect. In general, it is negligible.

Thus far, we have only considered solutions where the compensation state of the lunar crust is constant with time and independent of the position on the surface of the Moon. More complicated histories of the strength of the lithosphere might be possible, but a full parameter-space survey is beyond the scope of this work. In Extended Data Fig. 10m–o, we present three example TPW tracks for cases for which the compensation state varies monotonically with time.

Models that allow for the compensation state to increase with time (Extended Data Fig. 10n, o) produce TPW tracks that would markedly reduce the age of the epithermal neutron palaeopole (to only about 1.5 Gyr in Extended Data Fig. 10o). Although this sort of weakening of the lithosphere with time might not be physical, it is possible that the loading and isostatic adjustment (or non-adjustment) of the PKT mare basalts and other near-surface mass anomalies could replicate the effect of this time-varying compensation state. Thus, further study of the geologic and geophysical history of the PKT could provide insight into the long-term stability of lunar polar ice.

**Plausible modes of long-term hydrogen stability.** Our nominal TPW models suggest a source for the observed off-polar hydrogen (plausibly in the form of water ice) that predates the migration of the lunar spin axis. Because this hydrogen would have to survive for what could amount to several billion years, it may have experienced temperature conditions warmer than present. Therefore, it is important to consider the long-term stability of polar hydrogen.

Water ice will be stable if the temperatures in the first few metres of the Moon's surface remain below about 145 K. Even near the poles, directly illuminated surfaces will experience maximum temperatures that exceed 145 K, which leads to ground ice being stable only within polar craters or regions with high topographic relief[5]. Above this temperature, water ice (thicker than a single surface-bound monolayer) will sublime on geologic timescales, with rates exceeding 1 mm Gyr$^{-1}$ (refs 25, 81).

Although a single monolayer of water is more stable, it is probably not sufficient to cause the observed hydrogen excess. A typical sample of Apollo lunar regolith has a surface area of about 0.5 m$^2$ g$^{-1}$ (refs 82, 83). An idealized monolayer contains approximately $10^{15}$ molecules per square centimetre, so a monolayer contains approximately $5 \times 10^{18}$ molecules per gram of regolith. This corresponds to a mass of $1.5 \times 10^{-4}$ grams of $H_2O$, or 17 p.p.m. of hydrogen atoms. Although variations in grain size may change the ratio of surface area to volume (and thus the mass fraction of hydrogen), with these assumptions, adsorption of water molecules directly to regolith can contribute only a small fraction of the minimum plausible hydrogen concentration observed at the epithermal neutron palaeopole (Fig. 1a, b). Thus, we assume that the observed hydrogen corresponds to either water ice mixed within regolith (pore ice), or hydrogen bound within mineral grains.

Ancient ice must also survive billions of years of impact gardening, a process that will slowly mix ice with the surrounding regolith[13,84]. Impact gardening can result in both ice loss, because ice is brought to the warmer near-surface, and preservation of ice, because ice is buried under layers of protective, thermally insulating regolith. However, impact gardening processes will dominate only if the water ice is completely immobile (as would be the case for adsorbed water). Given even short windows of time with temperatures above about 70–90 K, buried water molecules can migrate towards the surface, driven solely by the water vapour concentration gradient between the regolith and the vacuum of space. Water will migrate upward until it hits the predicted ice stability depths (Fig. 1c, d, Extended Data Fig. 4) where it will remain and concentrate, because loss rates to space are slow enough (1 mm Gyr$^{-1}$) that it will not thermally sublimate over geologic time. The ice may be buried again, mixed into the regolith or lost by an impact related process, but, assuming some small amount of thermal mobility, it will again return to the ice stability depth. Therefore, even accounting for impact gardening, as long as temperatures remain greater than about 70–90 K, but never exceed about 145 K, the predicted depths should be a good proxy for detectable hydrogen.

Ice stability models presented here show that ice can be stable both at the current and proposed palaeopole orientations. Extended Data Figure 4e, f shows that there are large areas that are stable in the upper 2.5 m for ice both at the current lunar pole position and at the proposed palaeopole (ice stability depth is assigned as an average of the two models). However, if wander led to a spin pole much further from the palaeopole, ground ice would no longer be stable in these locations. To estimate how far a shadowed crater could move from the pole and still retain large amounts of ground ice, we look at previous studies examining the effects of lunar obliquity on ice stability. In such studies, a polar crater (Shackleton) was found to retain stable ice until the Moon tilted by >12° (refs 4, 49). We use this 12° limit as an approximate estimate for the maximum extent of polar wander that can occur with respect to a palaeopole and still allow for the preservation of water ice at the pole. In fact, some wander past the current pole would aid in the migration of buried water to the surface by creating slightly warmer conditions than present. At present, some cold traps are so cold (maximum temperatures $< 90$ K) that ice is effectively immobile[25], leaving it to be slowly buried by impact gardening[13]. TPW might have caused these areas to experience conditions warm enough that ice buried by impact gardening would migrate towards the surface ($\sim 90\,\mathrm{K} < T < 145\,\mathrm{K}$), driven by the concentration gradient (with the vacuum of space).

Although many of our TPW paths remain within this 12° ice-stability limit, suggesting that the hydrogen observed at the epithermal neutron palaeopole is plausibly water ice (Fig. 4a, b, Extended Data Fig. 10a–d, n, o), many do not. In these 'large wander' cases, the shadowed regions near the epithermal palaeopole may have experienced temperatures that exceeded the 145-K stability limit for water ice. This would suggest that the epithermal neutrons may be mineralogically trapped or bound hydrogen, rather than pore ice. Most mineralogies will have higher bonding strength than that of water to water, and be more stable to large temperature fluctuations. It may be that pore ice was originally stable at these locations, but has since been partially lost (perhaps via hydrothermally interacting with the surrounding regolith), leaving only the most stable forms of hydrogen behind. However, impact gardening will slowly bury the grains this water is bound to and thus limit the length of time such hydrogen will be in sufficient abundances to be detectable via neutron spectrometry.

It is also plausible that the observed hydrogen might never have been water. Hydrogen can implant into permanently shadowed regions both from Earth's magnetotail and by backscattering of solar-wind hydrogen off of nearby irradiated crater walls. However, an explanation of why such a mechanism would result in the observed antipodal ice distribution has not been proposed. Perhaps areas that once harboured water ice are more accepting of solar-wind hydrogen. The time required to build up about 100 p.p.m. of rim-entrapped hydrogen in a permanently shadowed region has been estimated[28] to be of the order of 200 Myr. If not continuously resupplied, then hydrogen trapped at defects in grain rims has a chance to escape, with this chance depending primarily on two variables: diffusion activation energy and temperature. A range of realistic activation energies were found[85] for which hydrogen would be retained at low lunar temperatures and for billions of years.

Regardless of the hydrogen source, defects in weathered grain rims can create a large volume for hydrogen retention, and may be sufficient to explain the observed hydrogen concentrations. The possible retention of hydrogen trapped at defects within the rims of the lunar grains themselves has been calculated[28,86]. The maximum concentration is set by the maximum retention of implanted hydrogen in laboratory experiments, and is about $2 \times 10^{17}\,\mathrm{cm}^{-2}$. The thickness of the implantation rim is taken to be 100 nm. For a lunar soil surface area of $0.5\,\mathrm{m^2\,g^{-1}}$, the maximum trapped-hydrogen concentration is about 1,700 p.p.m. (ref. 28).

Large-scale defects, such as radiation tracks, can react with water[87–91]. Such a situation could occur as once-stable ice deposits begin to sublimate. This reaction was shown to increase the specific surface area and porosity of lunar fines and retain water. However, it is unclear how long water might stay within the grain lattice once it is established there. We adopt a silicate lattice of $3.3\,\mathrm{g\,cm^{-3}}$ and, hence, an atomic density of about $10^{23}\,\mathrm{cm}^{-3}$; we assume a water density of about $10^{20}\,\mathrm{cm}^{-3}$. By using Fick's law[92], with a typical silicate diffusion coefficient of $D = 10^{-25}\,\mathrm{cm^2\,s^{-1}}$, and calculating the flux $J$ across a 1-μm lattice layer into the vacuum (so that $10^{20}\,\mathrm{cm}^{-3}$ drops to $0\,\mathrm{cm}^{-3}$ in $10^{-4}\,\mathrm{cm}$), we obtain

$$J = D\frac{\partial n}{\partial x} = 10^{-25}\,\mathrm{cm\,s^{-1}} \times \frac{10^{20}\,\mathrm{cm^{-3}}}{10^{-4}\,\mathrm{cm}} = 0.1\,\mathrm{cm^{-2}\,s^{-1}}$$

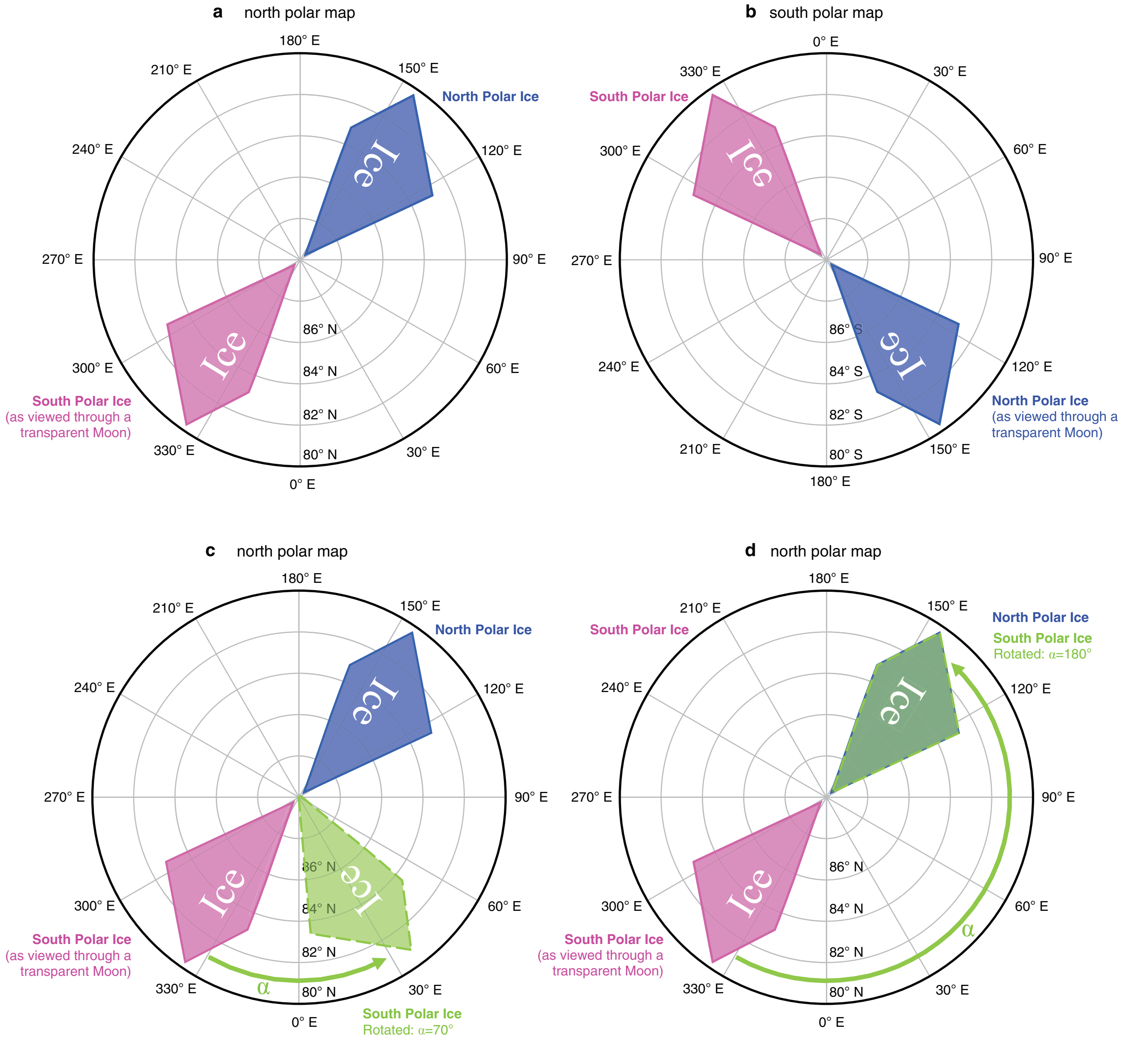
The number of water molecules in a $1\,\mu\mathrm{m} \times 1\,\mathrm{cm}^2$ volume is $10^{16}$, so the diffusion timescale is $10^{16}/(0.1\,\mathrm{s}^{-1}) = 10^{17}\,\mathrm{s} = 3\,\mathrm{Gyr}$. Using values from the literature[93,94], we estimate $D \approx 10^{-28}$ and $J \approx 10^{-4}\,\mathrm{cm^{-2}\,s^{-1}}$, which implies that water, having incorporated into lunar materials, will not diffuse from the outer micrometres in several billion years.

The mechanisms described above may allow for reasonable long-term (Gyr) storage of hydrogen—either in the form of pore ice or mineralogically bound hydrogen—in the off-polar regions detected by the epithermal neutron data presented here. However, the evidence presented here points to a correlation with preferential water stability along the path of TPW.

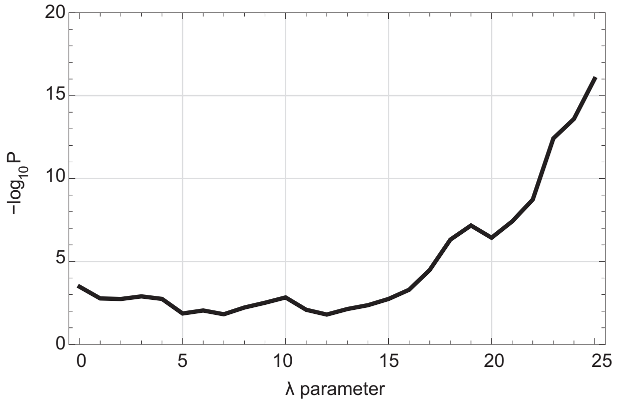
It is possible that the epithermal neutron distribution marks the surviving hydrogen from an epoch of ice stability or high supply (for example, the late heavy bombardment, a time during which the Moon had a protective magnetic field, or primordial internal water from the Moon's formation); alternatively, it could trace a history of later-stage addition of hydrogen (for example, outgassing of water from mare volcanism, large volatile-rich impacts or some variation in solar wind). Future orbital missions with high-resolution, high-precision neutron spectrometers might be able to better constrain the extent of the polar hydrogen, and future *in situ* polar landers or sample return might be able to directly determine the nature of lunar polar hydrogen. Current evidence, such as the detection of water ice in the LCROSS-impact vapour plume[26] (this impact occurred very close to our proposed southern palaeopole), suggests that the observed hydrogen enhancement is due to water, and that the Moon may not have wandered an extreme amount since the deposition of this water.

31. Mitrofanov, I. G. *et al.* Experiment LEND of the NASA Lunar Reconnaissance Orbiter for high-resolution mapping of neutron emission of the Moon. *Astrobiology* **8,** 793–804 (2008).
32. Litvak, M. L. *et al.* Global maps of lunar neutron fluxes from the LEND instrument. *J. Geophys. Res. Planets* **117,** E00H22 (2012).
33. Lawrence, D. J., Elphic, R. C., Feldman, W. C., Funsten, H. O. & Prettyman, T. H. Performance of orbital neutron instruments for spatially resolved hydrogen measurements of airless planetary bodies. *Astrobiology* **10,** 183–200 (2010).
34. Lawrence, D. J. *et al.* Technical Comment on "Hydrogen mapping of the lunar south pole using the LRO neutron detector experiment LEND". *Science* **334,** 1058 (2011).
35. Eke, V. R., Teodoro, L. F. A., Lawrence, D. J., Elphic, R. C. & Feldman, W. C. A quantitative comparison of lunar orbital neutron data. *Astrophys. J.* **747,** 6 (2012).
36. Teodoro, L. F. A., Eke, V. R., Elphic, R. C., Feldman, W. C. & Lawrence, D. J. How well do we know the polar hydrogen distribution on the Moon? *J. Geophys. Res. Planets* **119,** 574–593 (2014).
37. Feldman, W. C. *et al.* Gamma-ray, neutron, and alpha-particle spectrometers for the Lunar Prospector mission. *J. Geophys. Res. Planets* **109,** E07S06 (2004).
38. Maurice, S., Lawrence, D. J., Feldman, W. C., Elphic, R. C. & Gasnault, O. Reduction of neutron data from Lunar Prospector. *J. Geophys. Res. Planets* **109,** E07S04 (2004).
39. Lawrence, D. J. *et al.* Improved modeling of Lunar Prospector neutron spectrometer data: implications for hydrogen deposits at the lunar poles. *J. Geophys. Res.* **111,** E08001 (2006).
40. Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing* Ch. 14.5, 636–639 (1992).
41. Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an infinitely large population. *Biometrika* **10,** 507–521 (1915).
42. Watson, K., Murray, B. & Brown, H. On the possible presence of ice on the Moon. *J. Geophys. Res.* **66,** 1598–1600 (1961).
43. Harmon, J. K., Slade, M. A. & Rice, M. S. Radar imagery of Mercury's putative polar ice: 1999–2005 Arecibo results. *Icarus* **211,** 37–50 (2011).
44. Chabot, N. L. *et al.* Areas of permanent shadow in Mercury's south polar region ascertained by MESSENGER orbital imaging. *Geophys. Res. Lett.* **39,** L09204 (2012).
45. Lawrence, D. J. *et al.* Evidence for water ice near Mercury's north pole from MESSENGER neutron spectrometer measurements. *Science* **339,** 292–296 (2013).
46. Lawrence, D. J., Miller, R. S., Ozimek, M. T., Peplowski, P. N. & Scott, C. J. High-resolution mapping of lunar polar hydrogen with a low-resource orbital mission. *Acta Astronaut.* **115,** 452–462 (2015).
47. Tye, A. R. *et al.* The age of lunar south circumpolar craters Haworth, Shoemaker, Faustini, and Shackleton: implications for regional geology, surface processes, and volatile sequestration. *Icarus* **255,** 70–77 (2015).
48. Ward, W. R. Past orientation of the lunar spin axis. *Science* **189,** 377–379 (1975).
49. Siegler, M. A., Bills, B. G. & Paige, D. A. Effects of orbital evolution on lunar ice stability. *J. Geophys. Res. Planets* **116,** E03010 (2011).
50. Lambeck, K. *The Earth's Variable Rotation: Geophysical Causes and Consequences* (Cambridge Univ. Press, 1980).
51. Sabadini, R. & Vermeersen, B. *Global Dynamics of the Earth: Applications of Normal Mode Relaxation Theory to Solid-Earth Geophysics* (Kluwer, 2004).
52. Van Hoolst, T. in *Treatise on Geophysics: Planets and Moons* (eds Schubert, G. & Spohn, T.) 123–164 (Elsevier, 2009).
53. Colombo, G. in *Measure of the Moon* (eds Kopal, Z. & Goudas, C. L.) 12–22 (Springer, 1967).
54. Peale, S. J. Generalized Cassini's laws. *Astron. J.* **74,** 483–488 (1969).
55. Gold, T. Instability of the Earth's axis of rotation. *Nature* **175,** 526–529 (1955).
56. Goldreich, P. & Toomre, A. Some remarks on polar wandering. *J. Geophys. Res.* **74,** 2555–2567 (1969).
57. Matsuyama, I. Fossil figure contribution to the lunar figure. *Icarus* **222,** 411–414 (2013).
58. Melosh, H. J. Mascons and the Moon's orientation. *Earth Planet. Sci. Lett.* **25,** 322–326 (1975).
59. Runcorn, S. K. Lunar magnetism, polar displacements and primeval satellites in the Earth–Moon system. *Nature* **304,** 589–596 (1983).
60. Runcorn, S. K. The primeval axis of rotation of the Moon. *Phil. Trans. R. Soc. Lond. A* **313,** 77–83 (1984).
61. Hood, L. L. Central magnetic anomalies of Nectarian-aged lunar impact basins: probable evidence for an early core dynamo. *Icarus* **211,** 1109–1128 (2011).
62. Cournède, C., Gattacceca, J. & Rochette, P. Magnetic study of large Apollo samples: possible evidence for an ancient centered dipolar field on the Moon. *Earth Planet. Sci. Lett.* **331–332,** 31–42 (2012).
63. Tsunakawa, H., Takahashi, F., Shimizu, H., Shibuya, H. & Matsushima, M. Surface vector mapping of magnetic anomalies over the Moon using Kaguya and Lunar Prospector observations. *J. Geophys. Res. Planets* **120,** 1160–1185 (2015).
64. Kim, H. R., Hood, L. L., von Frese, R. R. B. & O'Reilly, B. E. Nectarian paleomagnetic pole inferred from Kaguya satellite magnetic observations of the central Leibnitz basin. *Lunar Planet. Sci. Conf.* **46,** 1914 (2015).
65. Iess, L. *et al.* The gravity field and interior structure of Enceladus. *Science* **344,** 78–80 (2014).
66. Matsuyama, I. & Nimmo, F. Tectonic patterns on reoriented and despun planetary bodies. *Icarus* **195,** 459–473 (2008).
67. Schenk, P., Matsuyama, I. & Nimmo, F. True polar wander on Europa from global-scale small-circle depressions. *Nature* **453,** 368–371 (2008).
68. Nimmo, F. & Manga, M. in *Europa* (eds Pappalardo, R. T. et al.) 381–404 (Univ. Arizona Press, 2009).
69. Perron, J. T., Mitrovica, J. X., Manga, M., Matsuyama, I. & Richards, M. A. Evidence for an ancient Martian ocean in the topography of deformed shorelines. *Nature* **447,** 840–843 (2007).
70. Kite, E. S., Matsuyama, I., Manga, M., Perron, J. T. & Mitrovica, J. X. True Polar Wander driven by late-stage volcanism and the distribution of paleopolar deposits on Mars. *Earth Planet. Sci. Lett.* **280,** 254–267 (2009).
71. Weiss, B. P. & Tikoo, S. M. The lunar dynamo. *Science* **346,** 1246753 (2014).
72. Bills, B. G. & Rubincam, D. P. Constraints on density models from radial moments: applications to Earth, Moon, and Mars. *J. Geophys. Res. Planets* **100,** 26305–26315 (1995).
73. Konopliv, A. S. *et al.* Improved gravity field of the Moon from Lunar Prospector. *Science* **281,** 1476–1480 (1998).
74. Zuber, M. T. *et al.* Gravity field of the Moon from the Gravity Recovery and Interior Laboratory (GRAIL) mission. *Science* **339,** 668–671 (2013).
75. Melosh, H. J. Large impact craters and the Moon's orientation. *Earth Planet. Sci. Lett.* **26,** 353–360 (1975).
76. Turcotte, D. L. & Schubert, G. *Geodynamics* 2nd edn, Ch. 4, 132–194 (Cambridge Univ. Press, 2002).
77. Muller, P. M. & Sjogren, W. L. Mascons: lunar mass concentrations. *Science* **161,** 680–684 (1968).
78. Melosh, H. J. *et al.* The origin of lunar mascon basins. *Science* **340,** 1552–1555 (2013).
79. Kendall, J. D., Johnson, B. C., Bowling, T. J. & Melosh, H. J. Ejecta from south pole-Aitken basin-forming impact: dominant source of farside lunar highlands. *Lunar Planet. Sci. Conf.* **46,** 2765 (2015).

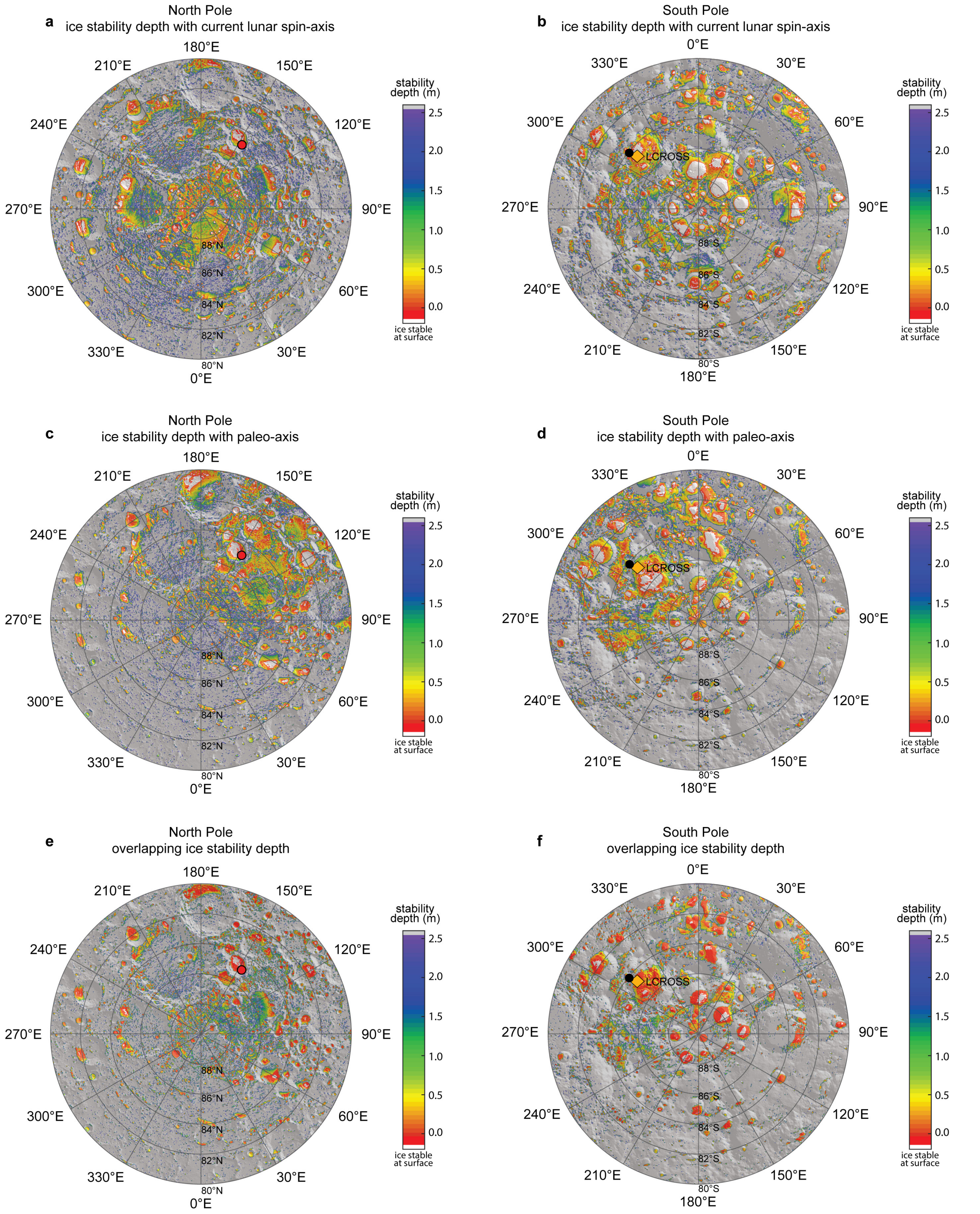
80. Arfken, G. B. & Weber, H. J. *Mathematical Methods for Physicists* 4th edn, 797–798 (Academic Press, 1995).
81. Schorghofer, N. The lifetime of ice on main belt asteroids. *Astrophys. J.* **682,** 697–705 (2008).
82. Cadenhead, D. A., Wagner, N. J., Jones, B. R. & Stetter, J. R. Some surface characteristics and gas interactions of Apollo 14 fines and rock fragments. *Proc. Lunar Planet. Sci. Conf.* **3,** 2243–2257 (1972).
83. Heiken, G. H., Vaniman, D. T., French, B. M. (eds) *Lunar Sourcebook: A User's Guide to the Moon* Ch. 7, 285–356 (Cambridge Univ. Press, 1991).
84. Crider, D. & Killen, R. M. Burial rate of Mercury's polar volatile deposits. *Geophys. Res. Lett.* **32,** L12201 (2005).
85. Farrell, W. M., Hurley, D. M. & Zimmerman, M. I. Solar wind implantation into lunar regolith: hydrogen retention in a surface with defects. *Icarus* **255,** 116–126 (2015).
86. Starukhina, L. V. Polar regions of the moon as a potential repository of solar-wind-implanted gases. *Adv. Space Res.* **37,** 50–58 (2006).
87. Fuller, E. L. Jr. Interaction of gases with lunar materials (12001): textural changes induced by sorbed water. *J. Colloid Interface Sci.* **55,** 358–369 (1976).
88. Gammage, R. B. & Holmes, H. F. Blocking of the water-lunar fines reaction by air and water concentration effects. *Proc. Lunar Sci. Conf.* **6,** 3305–3316 (1975).
89. Gammage, R. B. & Holmes, H. F. Alteration of Apollo 17 orange fines by adsorbed water vapor. *J. Colloid Interface Sci.* **55,** 243–251 (1976).
90. Gammage, R. B. & Holmes, H. F. Effect of annealing temperature on the reactivity of lunar fines toward adsorbed water. *Earth Planet. Sci. Lett.* **34,** 445–449 (1977).
91. Holmes, H. F. *et al.* Alteration of an annealed and irradiated lunar fines sample by adsorbed water. *Earth Planet. Sci. Lett.* **28,** 33–36 (1975).
92. Fick, A. V. On liquid diffusion. *Phil. Mag.* **10,** 30–39 (1855).
93. Brady, J. B. in *Mineral Physics & Crystallography: A Handbook of Physical Constants* (ed. Ahrens, T. J.) 269–290 (American Geophysical Union, 1995).
94. Zhang, Y. Diffusion in minerals and melts: theoretical background. *Rev. Mineral. Geochem.* **72,** 5–59 (2010).



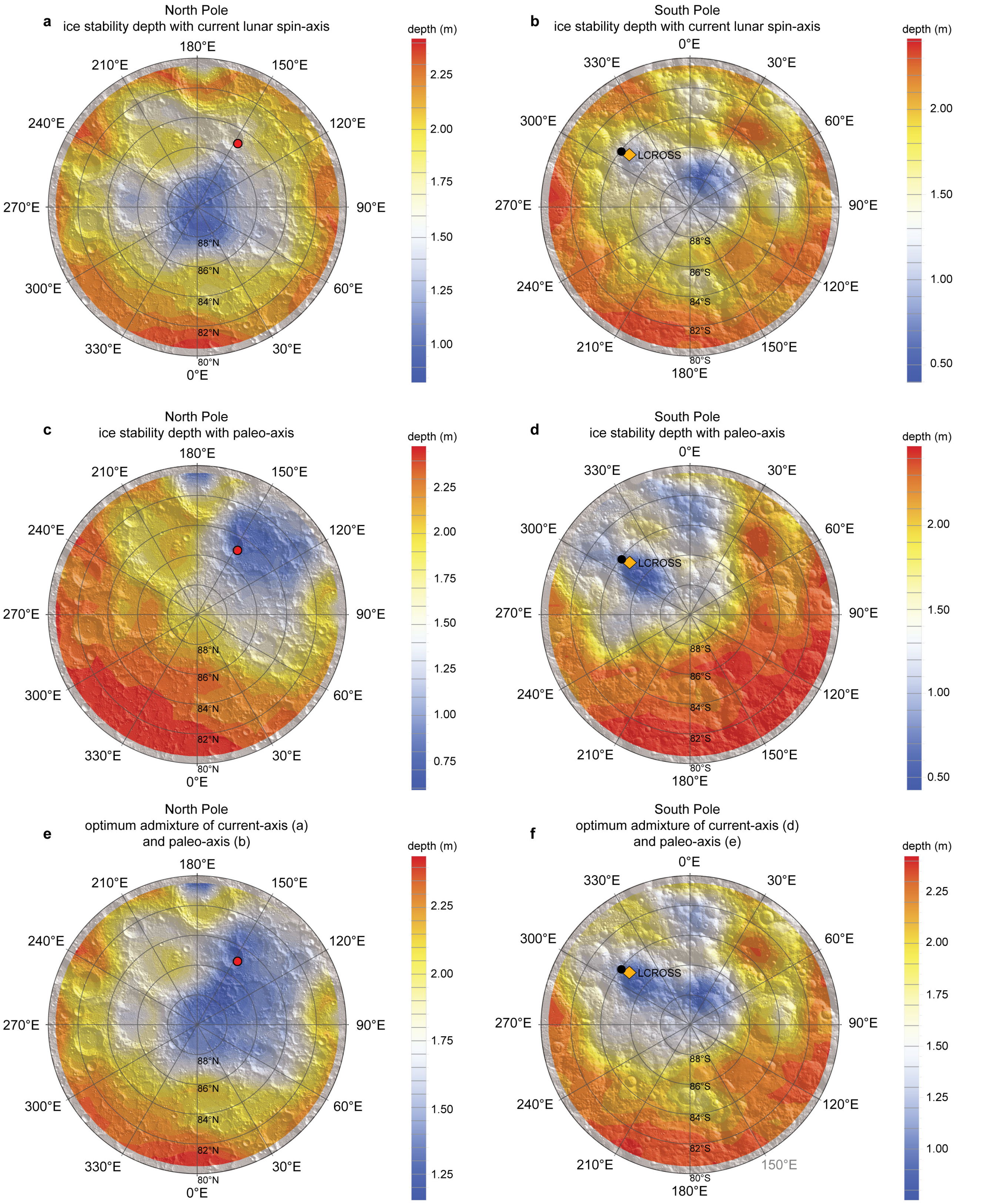
**Extended Data Figure 1 | Antipodal symmetry. a**, **b**, Standard north and south polar maps, respectively, akin to Fig. 1a, b. In each figure we show a hypothetical distribution of antipodally symmetric north (blue) and south (pink) polar ice. In each polar map, we show both the ice present in that hemisphere and the ice on the opposite hemisphere as it would be viewed through the Moon. Despite the antipodal symmetry, the polar ice that would be normally shown in each polar map (blue in **a**; pink in **b**) are not separated by 180° as one might expect. **c**, To test for antipodal symmetry, we rotate the features as viewed through the planet by an angle $\alpha$. **d**, Antipodally symmetric features match their antipodal counterparts if rotated by $\alpha = 180°$.



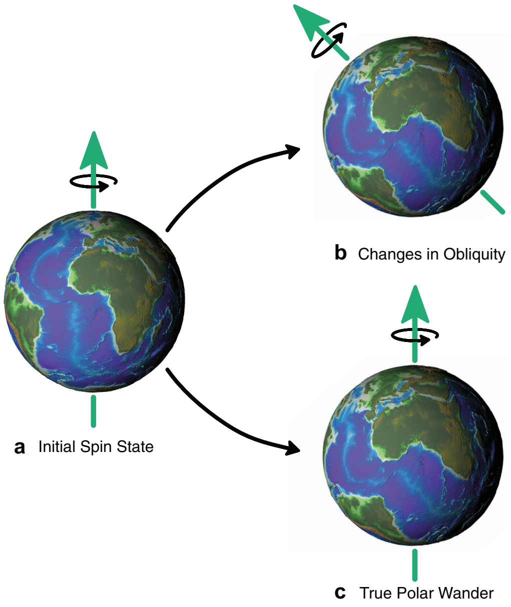
**Extended Data Figure 2 | Statistical significance of inter-polar hydrogen.** The statistical significant $-\log_{10}(P)$ is shown as a function of the likelihood parameter $\lambda$.


a
North Pole
ice stability depth with current lunar spin-axis
b
South Pole
ice stability depth with current lunar spin-axis
c
North Pole
ice stability depth with paleo-axis
d
South Pole
ice stability depth with paleo-axis
e
North Pole
overlapping ice stability depth
f
South Pole
overlapping ice stability depth
180°E
210°E
150°E
240°E
120°E
270°E
90°E
300°E
60°E
330°E
30°E
0°E
88°N
86°N
84°N
82°N
80°N
88°S
86°S
84°S
82°S
80°S
LCROSS
stability depth (m)
2.5
2.0
1.5
1.0
0.5
0.0
ice stable at surface

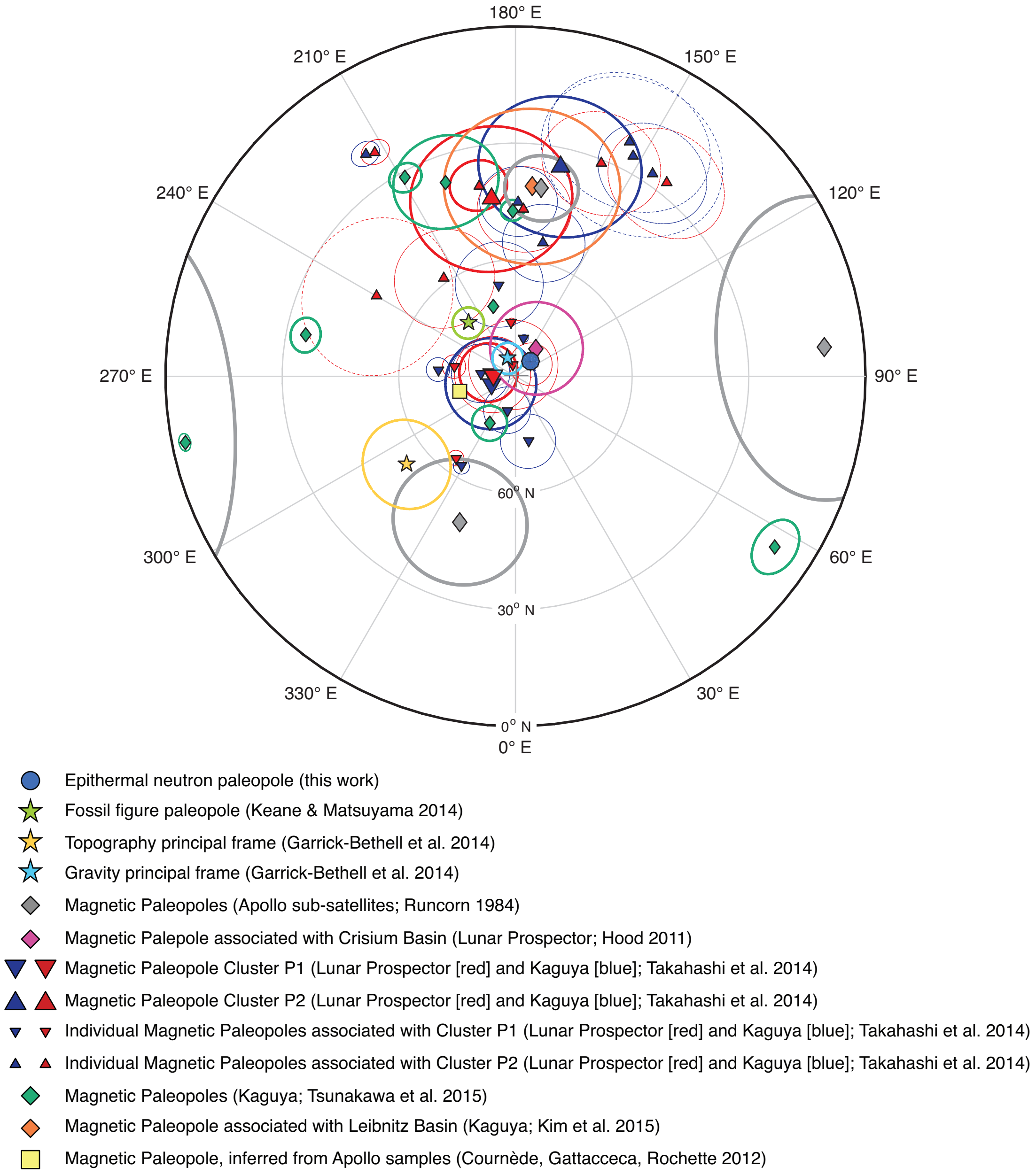
**Extended Data Figure 3 | Full-resolution maps of ice stability depth.** The ice stability depth is defined as the depth at which water ice will sublimate at a rate of 1 mm $\mathrm{Gyr}^{-1}$). **a**–**f**, Full-resolution maps of ice stability depth for areas where water ice is stable at the current orientation (**a**, **b**), in the palaeopole orientation (**c**, **d**) and in an 'overlapping' orientation (**e**, **f**); left panels shown the north polar region, right panels show the south. These models define the locations at which isotropically supplied water ice would be stable over geologic time; depth is given as an average of the two models in those locations. Panels **a**–**d** are the bases for Fig. 1c, d. Models are constrained to ± 300 km in stereographic *x*, *y* coordinates. Latitude lines are every 2° poleward of 80°. All other symbols are defined in Fig. 1. Panel **b** adapted from ref. 5, American Association for the Advancement of Science.



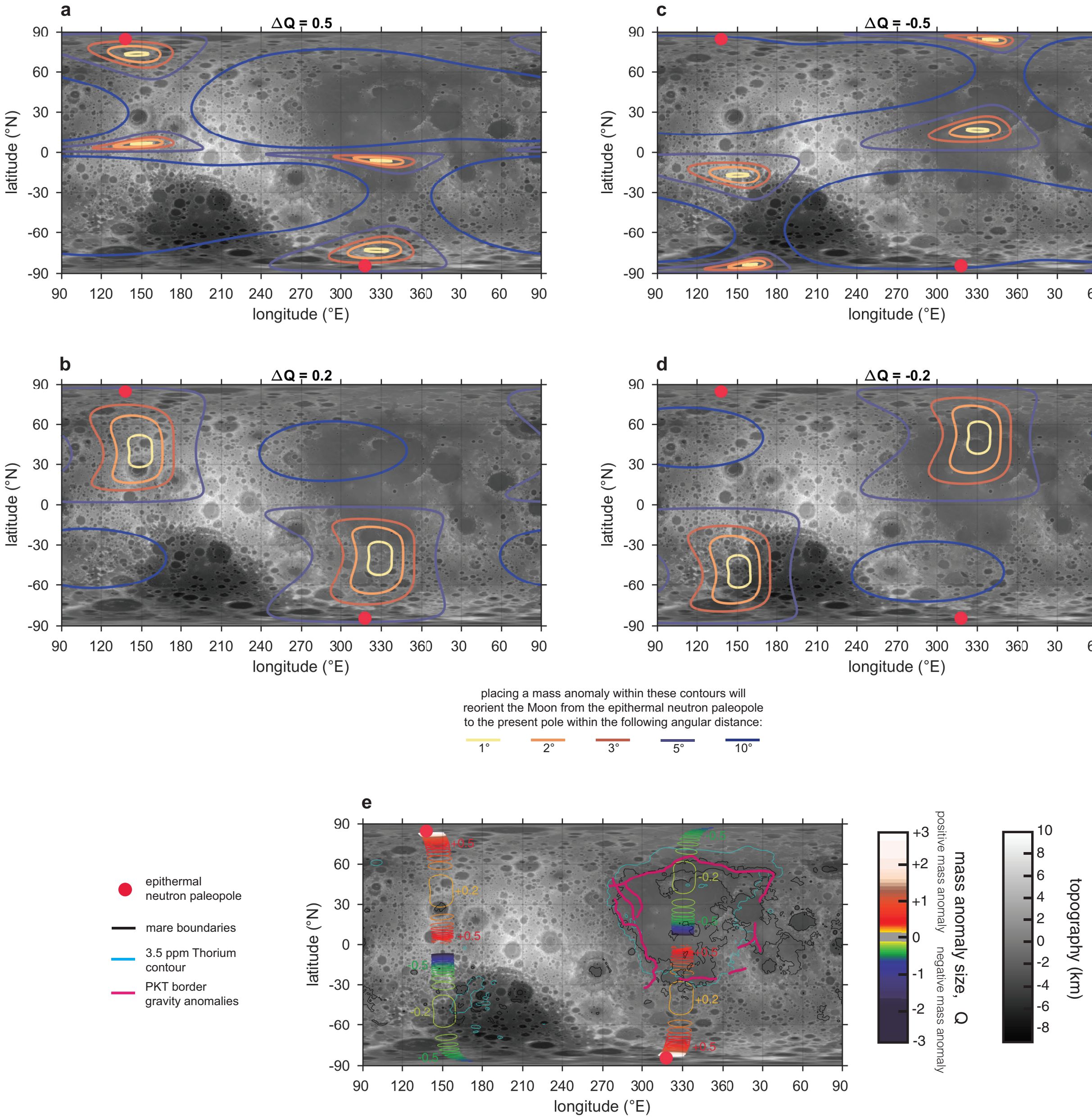
**Extended Data Figure 4 | Water-ice stability depths for the past and present poles. a–f,** Water-ice stability depths for the north (**a**, **c**, **e**) and south (**b**, **d**, **f**) polar regions. The model-derived stability depths[4–6] are shown for the current lunar spin axis (**a**, **b**) and the hypothesized palaeo-axis (**c**, **d**). Also shown is the optimum admixture of current- and palaeo-axis models that best matches the distributions of polar hydrogen abundance (**e**, **f**). Topography measured by the LOLA instrument[29] has been superimposed. Latitudinal contours are every 2° poleward of 80°. All other symbols are defined in Fig. 1.



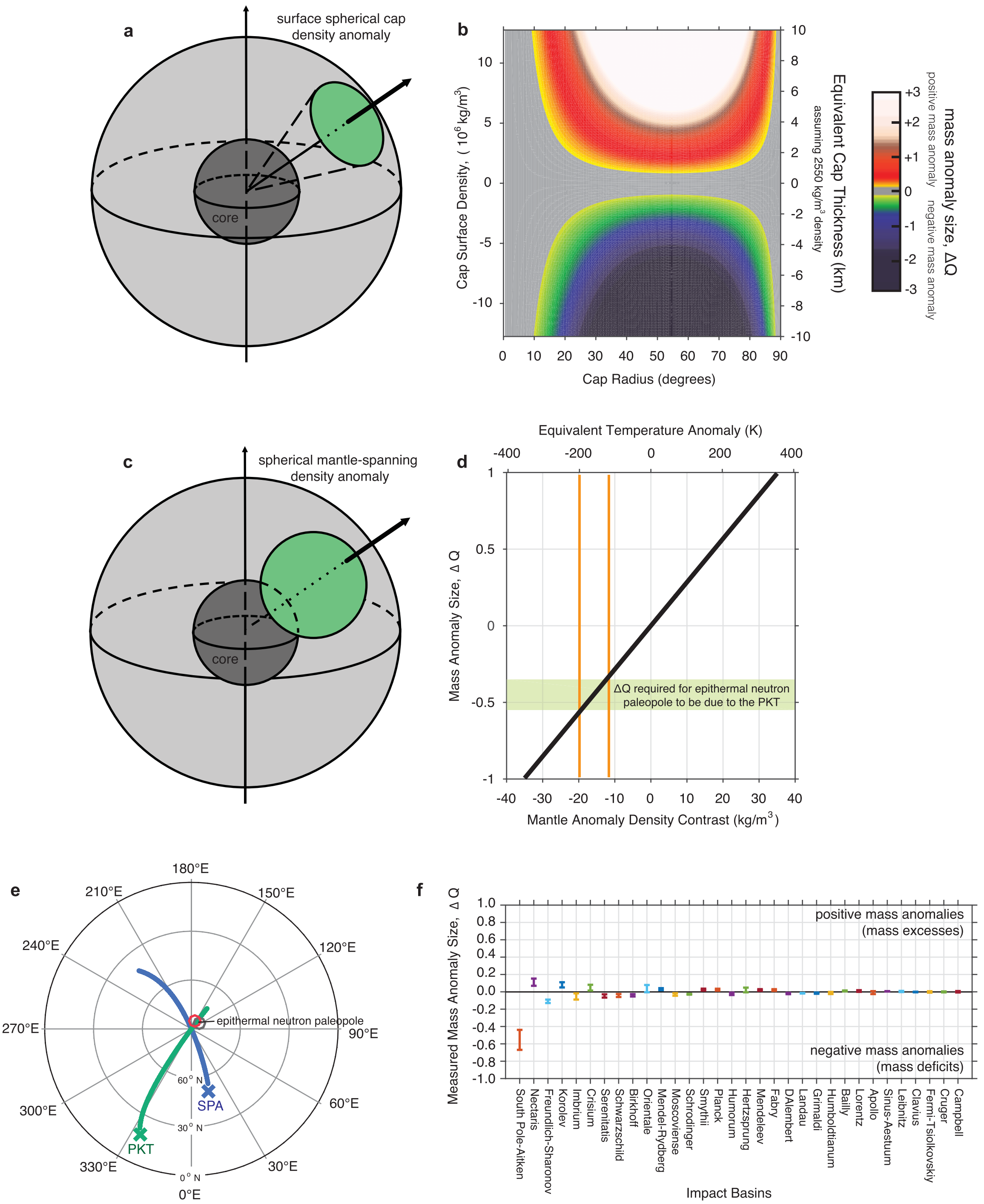
**Extended Data Figure 5 | Two different modes of planetary reorientation. a–c**, Initial spin state (**a**), changes in obliquity (**b**) and changes due to TPW (**c**). In these schematics, we view the reorienting planet in an inertial frame.



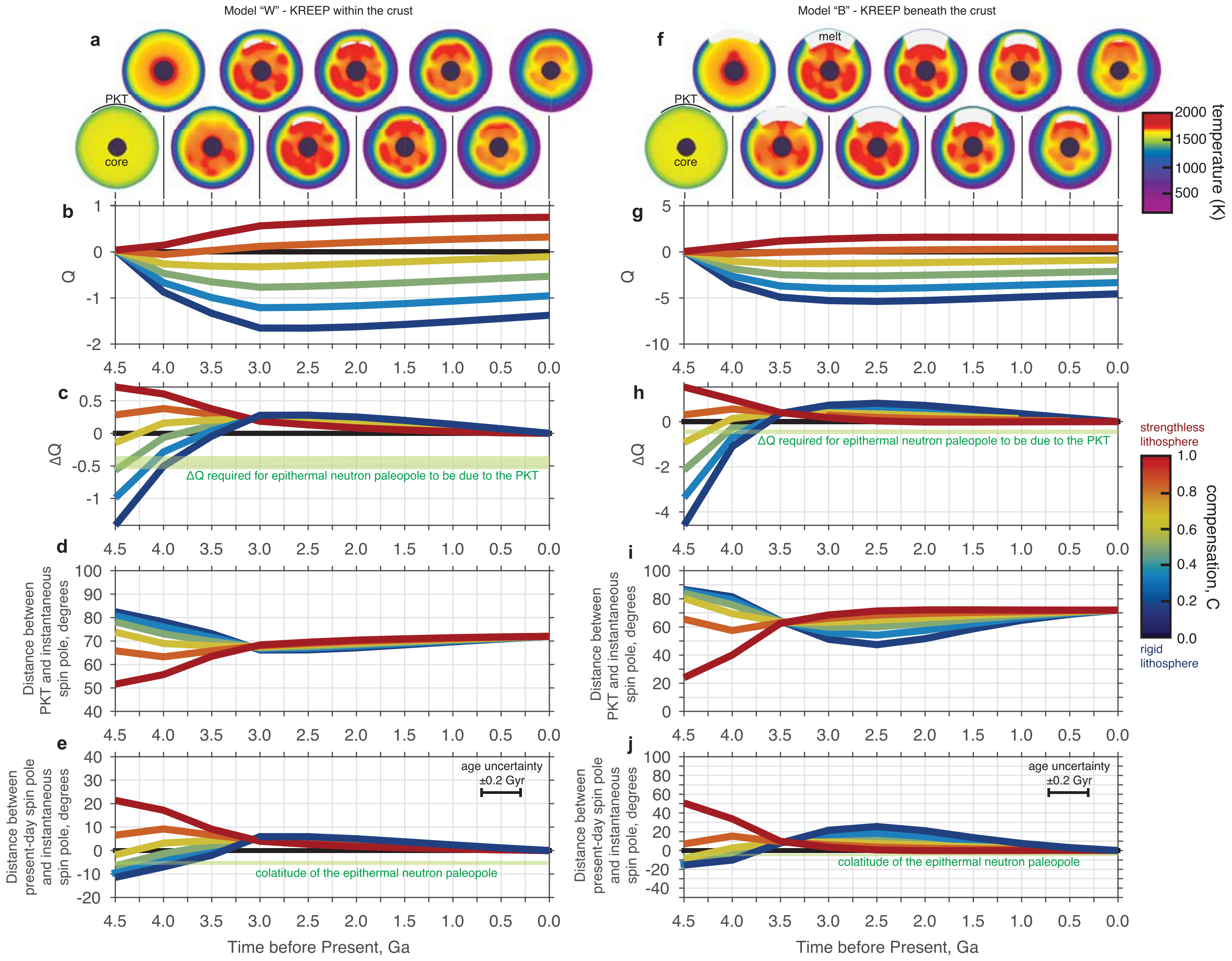
**Extended Data Figure 6 | The Moon's many published palaeopoles.** A collection of all published lunar palaeopoles, from a combination of palaeomagnetic data (diamonds[60,61,63,64], triangles[19] and square[62]), fossil figure estimates (stars[20,21]) and the epithermal neutron palaeopole reported here (circle). Ellipses around each point indicate 1$\sigma$ error.



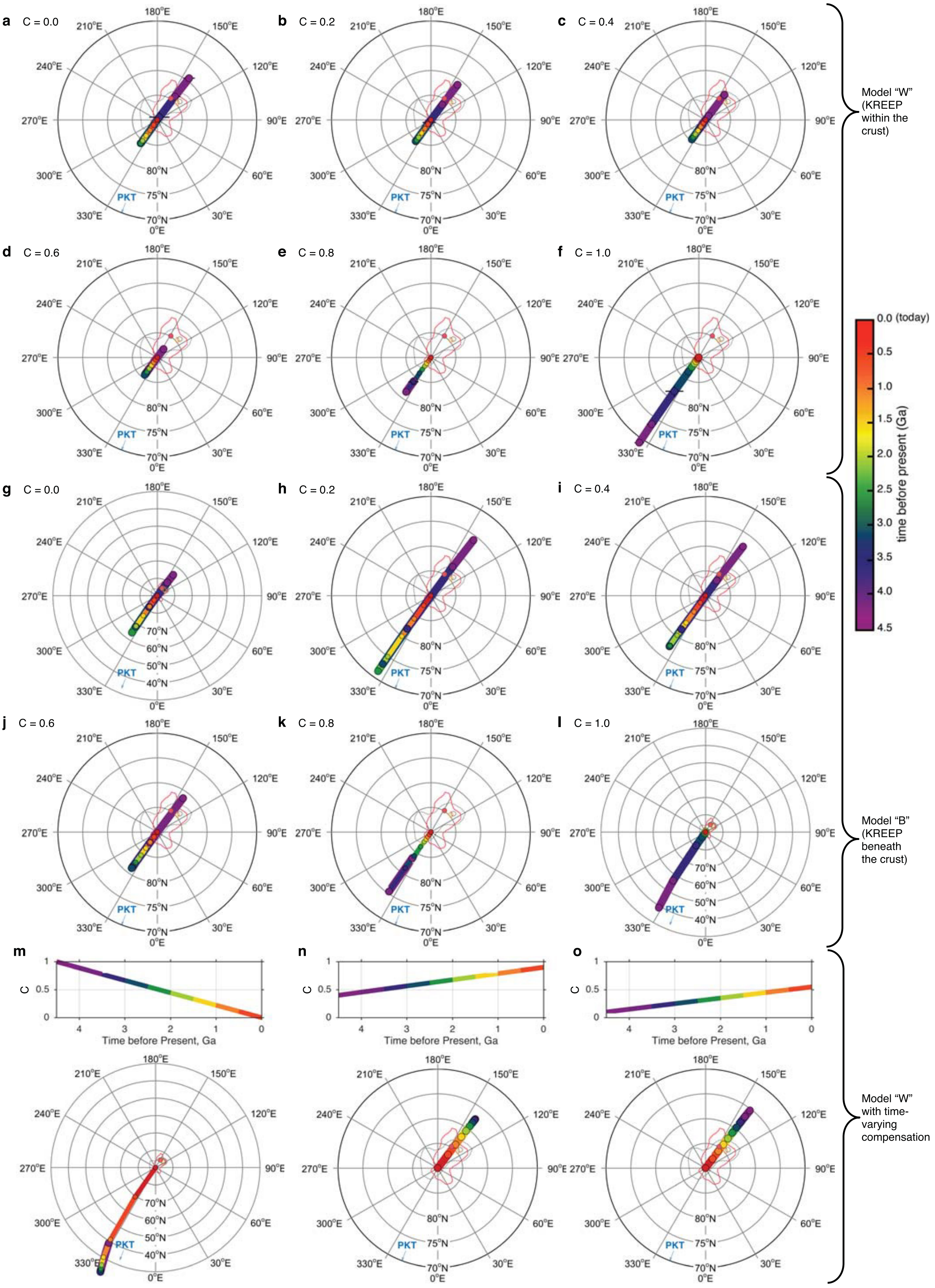
**Extended Data Figure 7 | Slices through the mass-anomaly parameter-space search. a–d**, Coloured contours enclose regions where placing a mass anomaly Δ$Q$ of the indicated size will cause a reorientation of the Moon to within the specified distance (see legend) of the present-day lunar spin pole. Red filled circles denote the epithermal neutron palaeopole. **e**, Contours enclose regions where mass anomalies Δ$Q$ must be centred to reorient the Moon from the epithermal neutron palaeopole to the present-day spin pole, to within 1°. This figure is the same as in Fig. 3a–c, but in an equirectangular projection; symbols and lines as in Fig. 3a–c.



**Extended Data Figure 8 | Simple physical models for reorienting the Moon, and the effect of lunar impact basins. a,** A schematic of our spherical cap model, showing a spherical cap on the surface of the Moon (green circle), centred on a particular latitude and longitude (arrow). **b**, The mass anomaly $\Delta Q$ for a spherical cap as a function of cap size and cap surface density (or cap thickness, assuming a density of $\rho = 2{,}550\,\mathrm{kg\,m^{-3}}$). **c**, A schematic of our mantle-spanning interior anomaly, with a spherical mass anomaly (green circle), centred on a particular latitude and longitude (arrow), grazing the core (dark grey circle) and the surface. **d**, The mass anomaly $\Delta Q$ for the spherical mantle anomaly as a function of the density contrast of the anomaly (black line). Over-plotted (green shading and orange lines) is the $\Delta Q$ required if the PKT is responsible ($\Delta Q = -0.45$; Fig. 3c). **e**, North polar projection of all possible palaeopole positions based on a mass anomaly placed at either the PKT or South Pole–Aitken basin (SPA). PKT paths always pass through the neutron palaeopole, whereas SPA paths are nearly orthogonal to this path. **f**, Mass anomalies $\Delta Q$ of the largest lunar impact basins derived from inverse fitting of the present-day lunar gravity field, following the method outlined in ref. 21. The only impact basin with a large enough mass anomaly is the South Pole–Aitken basin, which is not properly located to drive the observed TPW (Fig. 3b). The only major impact basin that is properly located is Moscoviense, which has a negligible mass anomaly (and with the wrong sign). Error bars are 1σ uncertainties from the inverse solution (see ref. 21).



**Extended Data Figure 9 | Mass anomalies and TPW of the Moon due the thermal evolution of the PKT. a–e**, Results for model W, in which KREEP is mixed within the crust. **f–j**, Results for model B, in which KREEP is mixed beneath the crust. **a**, **f**, Temperature cross-sections of the lunar mantle, as a function of time. The dark circle is the lunar core. White regions are partially molten. **b**, **c**, **g**, **h**, The mass anomaly $Q$ (**b**, **g**) and $\Delta Q$ (**c**, **h**) associated with the thermal anomalies shown in **a** and **f**, respectively, as functions of time for various assumed compensation states (coloured lines). $C = 0$ corresponds to a rigid lithosphere; $C = 1$ corresponds to a strengthless lithosphere. The region required to explain the epithermal neutron palaeopole is highlighted in green. **d**, **i**, The distance between PKT and the instantaneous spin pole as a function of time. **e**, **j**, The co-latitude of the instantaneous spin pole as a function of time. In this plot, the co-latitude is defined as positive if the northern spin pole is on the near side and negative if it is on the far side. The co-latitude of the epithermal neutron palaeopole is highlighted in green.



**Extended Data Figure 10 | Predicted TPW paths due to the thermal evolution of the PKT for a range of models and compensation states.** **a**–**f**, Model W, in which KREEP is mixed within the crust; **g**–**l**, model B, in which KREEP is mixed beneath the crust; **m**–**o**, model W, with a time-varying compensation state. In general, these TPW paths are consistent with the epithermal neutron pole forming early (4 ± 0.5 Gyr ago), as long as the lithosphere is partially rigid. If the lithosphere is weak or strengthless (**d**–**f**, **k**, **l**), the topographic uplift from the PKT thermal anomaly dominates and the TPW track never passes through the epithermal neutron pole. **m**, An example where the lithosphere starts strengthless (fluid; $C = 1$) and becomes perfectly rigid ($C = 0$) by the present day. **n**, **o**, Examples where the lithosphere starts partially rigid and becomes weaker with time. Although these cases may not be geophysically feasible, it is interesting as it confines the TPW paths to within the observed hydrogen distribution and reduces the age of the epithermal neutron palaeopole. Error bars indicate 1$\sigma$ uncertainty in the palaeopole position due to the rotational ambiguity of the PKT thermal models; the error bars are often smaller than the plotted pole positions. Contours and symbols as in Fig. 2a.

LETTER

doi:10.1038/nature16987

# Modes of surface premelting in colloidal crystals composed of attractive particles

Bo Li[1], Feng Wang[1], Di Zhou[1], Yi Peng[1], Ran Ni[2,3] & Yilong Han[1]

**Crystal surfaces typically melt into a thin liquid layer at temperatures slightly below the melting point of the crystal. Such surface premelting is prevalent in all classes of solids and is important in a variety of metallurgical, geological and meteorological phenomena[1]. Premelting has been studied using X-ray diffraction[2] and differential scanning calorimetry[3], but the lack of single-particle resolution makes it hard to elucidate the underlying mechanisms. Colloids are good model systems for studying phase transitions[4] because the thermal motions of individual micrometre-sized particles can be tracked directly using optical microscopy[5]. Here we use colloidal spheres with tunable attractions to form equilibrium crystal–vapour interfaces, and study their surface premelting behaviour at the single-particle level. We find that monolayer colloidal crystals exhibit incomplete premelting at their perimeter, with a constant liquid-layer thickness. In contrast, two- and three-layer crystals exhibit conventional complete melting, with the thickness of the surface liquid diverging as the melting point is approached. The microstructures of the surface liquids differ in certain aspects from what would be predicted by conventional premelting theories. Incomplete premelting in the monolayer crystals is triggered by a bulk isostructural solid–solid transition and truncated by a mechanical instability that separately induces homogeneous melting within the bulk. This finding is in contrast to the conventional assumption that two-dimensional crystals melt heterogeneously from their free surfaces[3,6] (that is, at the solid–vapour interface). The unexpected bulk melting that we observe for the monolayer crystals is accompanied by the formation of grain boundaries, which supports a previously proposed grain-boundary-mediated two-dimensional melting theory[7]. The observed interplay between surface premelting, bulk melting and solid–solid transitions challenges existing theories of surface premelting and two-dimensional melting.**

Surface premelting behaviour can be categorized into two types[1]: complete premelting, whereby the thickness of the liquid layer $l$ diverges as the melting point of the crystal is approached, and incomplete premelting, whereby $l$ increases as the melting point is approached, but remains finite. Incomplete premelting is attributed to the effects of the frequency-dependent dispersion force in Lifshitz theory[1], but this supposition is difficult to test. A special case of incomplete premelting is blocked surface premelting, which features a constant $l$ in the premelting temperature regime[8–10]; however, the underlying mechanism for this behaviour is unclear. The effect of dimensionality on surface premelting behaviour is also poorly understood. The surface plays a vital role in bulk phase transitions (for example, bulk melting starts from surfaces[6]), but little is known about the effects of the bulk on surface behaviour, which partly because the bulk and surface regions have rarely been measured simultaneously in atomic experiments and simulations.

Motivated by these open questions, we sought to observe both the surface and the bulk of colloidal crystals during premelting at the single-particle level. Colloids have provided insight into the microscopic kinetics of various bulk phase transitions, such as melting, crystallization, glass transition and solid–solid transition[4]. These transitions have usually been studied in repulsive colloids, which form solid phases only at high density under geometric confinement. However, equilibrium solid–vapour interfaces are formed only by attractive particles. Attractive colloidal particles better mimic atoms because all atoms have attractions and can condense into solid in vapour. To drive a thermodynamic phase transition in crystals, the colloids should be tunable. Most tunable attractions in colloids are either extremely short ranged[11] or difficult to tune precisely in a useful range of attraction strengths (that is, $0-1k_BT$, where $k_B$ is the Boltzmann constant and $T$ is the temperature)[12,13].

Here, we report relatively long-range dye-induced attractions between colloidal poly(methylmethacrylate) (PMMA) spheres with strengths that can be finely tuned over a temperature range of 20–30 °C (Fig. 1, Supplementary Figs 2 and 3). PMMA spheres with diameters of $\sigma = 2.02\,\mu$m were assembled into triangular lattices with millimetre-sized crystalline domains and confined in a parallel-plate glass cell. Premelting and melting were induced by lowering the temperature (that is, weakening the attraction). At each temperature step, we recorded the thermal motion of approximately 6,000 particles under equilibrium for about 1 h. Experimental details are provided in Supplementary Information.

The free surface of a solid premelts when $\Delta\gamma \equiv \gamma_{cv} - (\gamma_{cl} + \gamma_{lv}) > 0$, in which $\gamma$ is the interfacial energy, and the subscripts 'cv', 'cl' and 'lv'

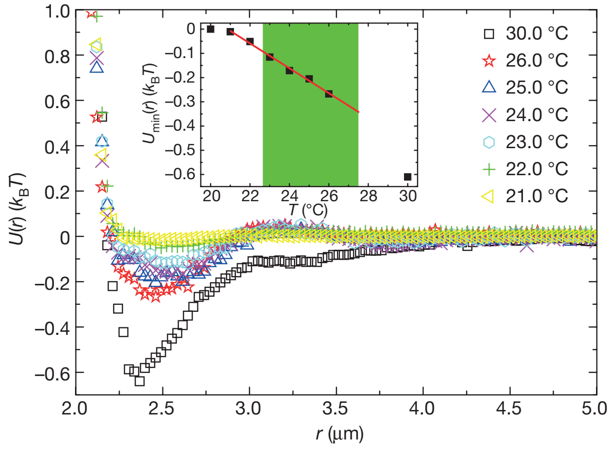


**Figure 1 | Pair potentials of the 2.02-μm-diameter PMMA spheres measured from monolayer liquid structures.** Pair potentials $U(r)$ are shown as a function of interparticle distance $r$ for various temperatures, as indicated. Inset, the attraction strength $U_{min}(r)$ decreases linearly (red line) in the green shaded temperature region, which corresponds to the temperatures that were used in the premelting experiment (Fig. 2). $U_{min}(r)$ was used to determine the effective temperature $k_BT/|U_{min}(r)|$, because the absolute temperature $T$ was almost constant during the premelting (approximately 300 K). See also Supplementary Fig. 1.

[1]Department of Physics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China. [2]School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore 637459, Singapore. [3]Van't Hoff Institute for Molecular Sciences, Universiteit van Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.

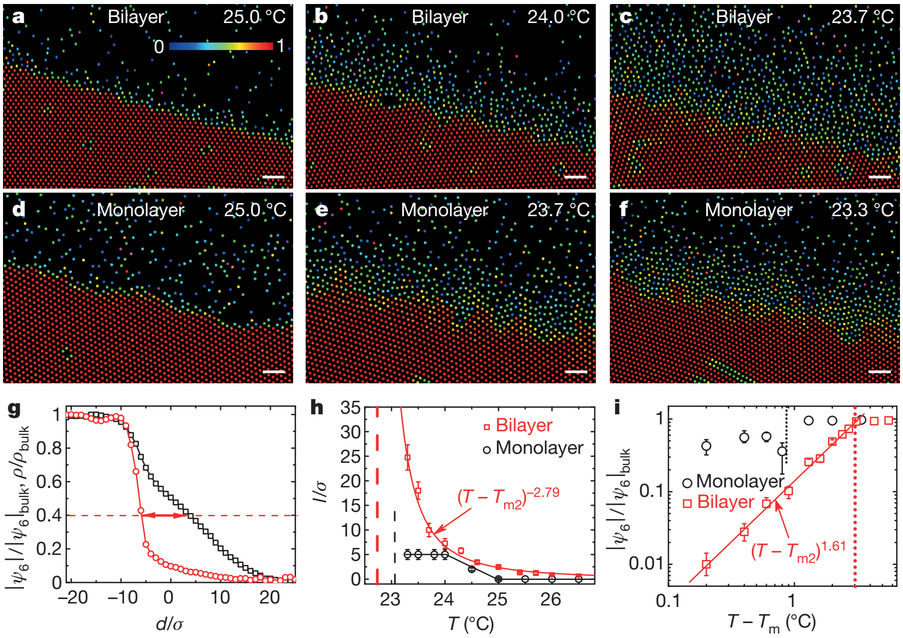


**Figure 2 | Surface premelting. a–f**, Surface premelting of bilayer (**a–c**; see also Supplementary Video 1) and monolayer (**d–f**; see also Supplementary Video 2) crystals. The colour scale (shown in **a**) indicates the magnitude of the orientational-order parameter of each particle, $|\psi_{6j}|$. Scale bars, 10 μm. **g**, Surface profiles of $|\psi_6|(d)$ (red) and $\rho(d)$ (black) for a bilayer crystal at 23.7 °C, normalized by their bulk values. The horizontal dashed line and arrow indicate the surface thickness $l$. **h**, As the temperature decreased, the thickness of the premelted liquid layer $l$ (normalized by the particle diameter $\sigma$) increased according to $l(T) \propto (T - T_{m2})^{-n}$ (with $T_{m2} = 22.7$ °C and $n = 2.79$) for the bilayer crystals, but saturated and remained constant for the monolayer crystals. The vertical thick (red) and thin (black) dashed lines represent the bulk melting points of the bilayer and monolayer crystals, respectively. **i**, The orientational-order parameter behaved according to $|\psi_6|_{d=0} \propto (T - T_{m2})^m$, with $T_{m2} = 22.7$ °C and $m = 1.61$, for the bilayer crystals ($d = 0$ is shown in Supplementary Fig. 10). The vertical thick (red) and thin (black) dotted lines represent the premelting points of the bilayer and monolayer crystals, respectively. Plots of $l/\sigma$ and $|\psi_6|_{d=0}$ as functions of the effective temperature $k_BT/U_{\min}$ are provided in Supplementary Fig. 13; similar power laws to those shown in **h** and **i** are obtained. We broke the data into three periods of time, and the errors in **h** and **i** are the standard deviations from the three subsets.

represent the crystal–vapour, crystal–liquid and liquid–vapour interfaces, respectively. We defined the thickness of the surface liquid $l$ as $|d_1 - d_2|$, with $d_{1,2}$ such that $|\psi_6|(d = d_1)/|\psi_6|_{\text{bulk}} = 0.4 = \rho(d = d_2)/\rho_{\text{bulk}}$ (Fig. 2g)[14], in which $|\psi_6|$ is the orientational order of the liquid, $\rho$ is its density, $d$ is the distance to the original (before premelting) solid–vapour interface, and 'bulk' superscripts indicate the corresponding bulk crystal values; we chose this definition because $|\psi_6|(d)$ and $\rho(d)$ should decrease at the crystal–liquid and liquid–vapour interfaces, respectively. The local density $\rho_j$ is the inverse area of the Voronoi cell of particle $j$. The local $\psi_{6j} = n_j^{-1}\sum_{k=1}^{n_j} e^{6i\theta_{jk}}$, in which $\theta_{jk}$ is the angle of the bond between particle $j$ and its neighbour $k$, and $n_j$ is the number of nearest neighbours of particle $j$ (ref. 15), with $0 \le |\psi_{6j}| \le 1$ and a higher $|\psi_{6j}|$ representing better six-fold symmetry. The expressions $|\psi_{6j}|(d)$ and $\rho(d)$ were derived by averaging $|\psi_{6j}|$ and $\rho_j$, respectively, over all particles at a distance $d$ from the original (before premelting) solid–vapour interface (Supplementary Figs 9–11). Landau theory predicts that both $|\psi_{6j}|(d)$ and $\rho(d)$ decay exponentially[16]; this prediction has been tested using simulations[17], but had not previously been tested experimentally. We found that $|\psi_{6j}|(d)$ decayed exponentially in bilayer crystals, but $\rho(d)$ did not (Fig. 2g and Supplementary Fig. 9). The decay length of $|\psi_{6j}|(d)$ was approximately three layers, which is in good agreement with results from simulations of premelted ice[18]. The linear decay of $\rho(d)$ illustrated in Fig. 2g and Supplementary Fig. 9 has not previously been observed experimentally or predicted by theory.

For the bilayer crystals, the liquid thickness was adequately fitted using $l(T) \propto (T - T_{m2})^{-n}$, in which $T_{m2} = 22.7$ °C is the melting temperature (Fig. 2h). Such power-law behaviour of $l(T)$ is typical for complete premelting, which has been predicted by theory[1] and simulation[19] and is commonly observed in metals[2]. By contrast, for the monolayer crystal, $l(T)$ abruptly increased at the onset of premelting (at a temperature $T_{pm1}$) and remained constant between melting and premelting (for $T_{m1} < T < T_{pm1}$), indicating a blocked premelting (Fig. 2d–f, h).

Landau theory predicts that the order parameter at the original position of the solid–vapour interface (that is, $d = 0$; see Supplementary Fig. 10) decreases according to $(T - T_m)^m$ rather than dropping directly to zero during complete premelting[20]; hence, some order is maintained in the surface liquid throughout premelting. The single-particle resolution we achieved is necessary to test this prediction, and enabled us to monitor the order parameter $|\psi_6|$ at a specific particle layer in the surface liquid (Supplementary Fig. 11); we confirmed that $|\psi_6|_{d=0} \propto (T - T_{m2})^m$ for the bilayers (Fig. 2i). Landau theory predicts that $m = n - 1$ (ref. 20), in which $n$ is the exponent in the expression for $l(T)$ (see Fig. 2h). Although this prediction is for 3D crystals, we observed that it was essentially applicable to our bilayer crystals, with the fitted $m = 1.61 \pm 0.06$ and $n = 2.79 \pm 0.08$.

At the onset of premelting $T_{pm1} = 23.9$ °C, we observed a lattice dilation in the bulk 2D crystal (Supplementary Fig. 14 and Supplementary Video 3) characterized by abrupt changes in density (Fig. 3a) and elastic moduli (Fig. 3b, c). The elastic moduli were measured from the dispersion relations of the lattice vibration calculated from displacement covariance matrices (Supplementary Fig. 18)[21,22]. The abrupt dilation of the lattice maintained the same lattice symmetry; hence, this dilation corresponded to an isostructural solid–solid phase transition. Isostructural solid–solid transitions generally occur in metallic and multiferroic systems[23], but have not been observed at the single particle level. The colloidal solid–solid transition was induced by the competition between the interaction energy and the free-volume entropy (Supplementary Fig. 15), and has been observed in simulations of attractive spheres in both 2D and 3D, but these simulations did not include surface effects[24]. We consider such a bulk solid–solid transition at $T_{pm1}$ to be the trigger of the surface premelting (that is, it is not a coincidence) because the expanded lattice is less attracted to surface particles and subsequently could trigger the premelting. The lattice dilation altered the interfacial energies and resulted in $\Delta\gamma > 0$, causing the formation of a surface liquid layer under thermal equilibrium. After the solid–solid transition, the lattice slightly expanded at $T_{m1} < T < T_{pm1}$ (Fig. 3a), which may have suppressed the surface liquid growth. In contrast to the monolayer crystals, no solid–solid transition was observed in the bilayer bulk crystals. The density and elastic moduli of the bulk crystals decreased continuously around the premelting point (Fig. 3a, d). Therefore, the thickness of the surface liquid increased less abruptly compared with that of the monolayer crystals. This is consistent with the conventional understanding that surface premelting is a precursor to bulk melting rather than it being a phase transition[1].

Bulk melting terminates the premelting process and is strongly dependent on dimensionality. 3D melting is a first-order phase

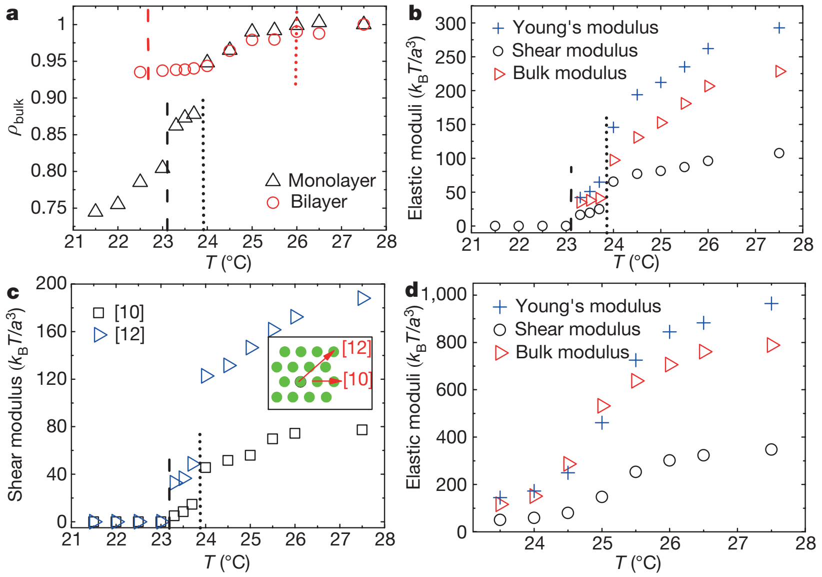


**Figure 3 | Density and elastic moduli. a**, Bulk densities $\rho_{\mathrm{bulk}}$ of monolayer and bilayer crystals, normalized by their corresponding values at 27.5 °C. **b**, Elastic moduli of the monolayer bulk crystal. The shear modulus was averaged over all directions. **c**, The shear modulus is lowest in the [10] lattice direction and highest in the [12] direction of the monolayer crystal (see inset). **d**, Elastic moduli of the bilayer crystal. The moduli in **b**–**d** are given in units of $k_BT/a^3$, with $a$ the lattice constant. The vertical dashed and dotted lines in **a**–**c** represent the melting and premelting temperatures, respectively.

transition[25], whereas 2D melting typically occurs via two continuous transitions through an intermediate hexatic phase as in the Kosterlitz–Thouless–Halperin–Nelson–Young (KTHNY) theory[15,26]. An alternative 2D melting theory predicts a one-step first-order transition that involves the formation of grain boundaries[7]. Most experiments and simulations have revealed the existence of a hexatic phase, although the transitions may be first-order or continuous for various repulsive particle interactions[27–29]. One-step first-order melting transitions have been observed in molecular monolayers without the observation of grain boundaries[15]. Grain-boundary-mediated melting was observed in a polycrystalline plasma monolayer[30], but the formation of grain boundaries in 2D single-crystal melting has yet to be directly observed.

We observed that the bilayer crystals underwent conventional surface melting[3] as a continuation of surface premelting; specifically, the premelted liquid propagated from the surface into the bulk during the melting process (Fig. 4a–c, Supplementary Video 4). By contrast, the monolayer crystals broke down into a liquid from both the surfaces and within the bulk (Fig. 4d–f, Supplementary Fig. 16 and Supplementary Video 5). Figure 4f, Supplementary Fig. 16d–f and Supplementary Video 6 illustrate either structural liquids with transient polycrystalline patches or a crystal–liquid coexistence, but not polycrystalline solids. Particles actively swapped positions with their neighbours near grain boundaries and occasionally inside crystalline patches. As the interparticle attraction decreased, more liquid particles with low $|\psi_6|$ were generated through the formation of new grain boundaries. Grain boundaries formed with an equal probability throughout the entire lattice (that is, a homogenous melting). This result contradicts the conventional assumption that 2D crystals heterogeneously melt from free surfaces[3,6]. According to 2D melting theory[7,15,26], when the dislocation core energy $E_c > 2.84k_BT$, the melting proceeds according to KTHNY theory with a proliferation of dislocations[7,15]. When $E_c < 2.84k_BT$, dislocations condense into strings as grain boundaries and the melting proceeds according to the grain-boundary-mediated melting theory[7]. We measured $E_c = 1.82k_BT$ at the melting point (Supplementary Fig. 17b)[31], which is consistent with the observed grain-boundary proliferation. Although both KTHNY theory and the grain-boundary-mediated melting theory apply only to infinitely large defect-free single crystals without free surfaces, we observed that grain-boundary-mediated melting occurred in 2D crystals with free surfaces. Below the melting point, the lattice expansion (Fig. 3a) generated vacancies (Supplementary Fig. 16b, c), which provided free space for forming grain boundaries. Vacancies typically do not form in the melting of repulsive 2D crystals, and the role of vacancy has not been thoroughly considered in 2D melting theories[7,15,26]. In addition, we determined that the shear modulus vanished in only the [10] direction at $T_{\mathrm{m1}}$ (Fig. 3c). This mechanical instability induced the bulk lattice to break down from within and

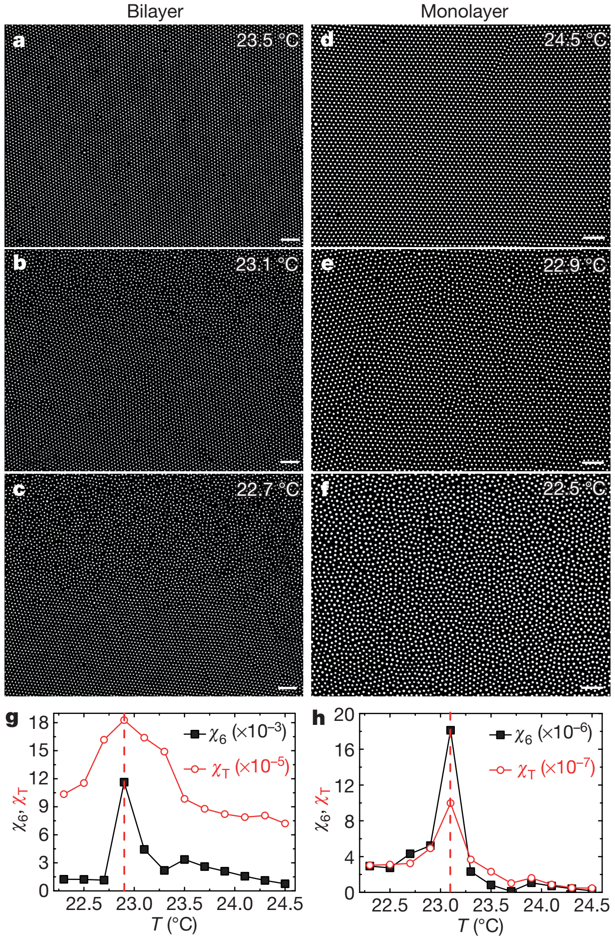


**Figure 4 | Bulk melting. a**–**f**, Raw images of the bulk melting of the bilayer (**a**–**c**; see also Supplementary Video 4) and monolayer (**d**–**f**; Supplementary Video 5) crystals. Scale bars, 10 μm. The image in **c** was captured during the melting process before the crystal completely melted, whereas the other images were captured at equilibrium. **g**, **h**, The orientational ($\chi_6$) and translational ($\chi_T$) susceptibilities of the bilayer crystal (**g**) and the monolayer crystal (**h**) peak at the bulk melting points (which are indicated by the vertical dashed lines). The peak values were measured from the bulk crystal without surface liquid.

resulted in a homogeneous melting, rather than a heterogeneous surface melting. The mechanical instability also interrupted the surface premelting, rendering it incomplete.

The melting point of the 2D crystals was identified from the density jump at 23.1 °C (Fig. 3a), which is a clear signature of a strong first-order transition. This melting point was confirmed from the peak positions of the orientational susceptibility $\chi_6 = \lim_{A\to\infty} A(\langle\psi_6^2\rangle - \langle\psi_6\rangle^2)$ and the translational susceptibility $\chi_\mathrm{T} = \lim_{A\to\infty} A(\langle\psi_\mathrm{T}^2\rangle - \langle\psi_\mathrm{T}\rangle^2)$ in Fig. 4h. (The global orientational-order parameter $\psi_6 = N^{-1}\sum_{j=1}^{N}\psi_{6j}$ for $N$ particles in area $A$, and the global translational-order parameter $\psi_\mathrm{T} = N^{-1}\sum_{j=1}^{N}\mathrm{e}^{i\boldsymbol{G}\cdot\boldsymbol{r}_j}$, in which $\boldsymbol{r}_j$ is the position of particle $j$ and $\boldsymbol{G}$ is the primary reciprocal-lattice vector[28].)

In contrast to the 2D crystals, identifying the melting point of the bilayer crystals by direct visualization was difficult because of the similarity in the appearance of a crystal exhibiting a thick premelted liquid layer before undergoing bulk melting and one exhibiting a possible solid–liquid coexistence after the bulk melting. We found that the melting point determined from the bulk susceptibilities peaks of 22.9 °C (Fig. 4g) is consistent with the fitted divergence point of $l$ at $T_{\mathrm{m}2} = 22.7$ °C (Fig. 2h), considering that our temperature step size was 0.2 °C and the uncertainty was 0.1 °C. The entire crystal completely melted at 22.7 °C. Bilayer crystals did not break down from within because of their stronger mechanical stability than monolayer crystals (Fig. 3). The finite shear moduli of the bilayer crystals at $T_{\mathrm{m}2}$ (Fig. 3d) indicated that their melting was induced by a thermodynamic instability rather than a mechanical instability.

We observed premelting, melting and solid–solid transition behaviour in 20 monolayer and 20 bilayer samples of spheres with diameters of 2.02 μm and 2.74 μm and obtained robust results. The attraction minimum was always 0.4 μm from the surface of the PMMA spheres; hence, larger spheres exhibited a shorter effective attraction range (Supplementary Fig. 2). The attraction range could be tuned by changing the size of the spheres, which can mimic a broad class of atoms. Different surface lattice orientations exhibited the same premelting behaviour with similar $l(T)$ values (Supplementary Video 2), indicating that the interfacial energy is not sensitive to the surface lattice orientation in colloids. The three-layer samples demonstrated similar complete premelting, but the thickness of the surface liquid increased more rapidly, owing a higher power-law exponent in the expression for $l(T)$ (Supplementary Fig. 12). These results suggest that dimensionality is a crucial factor for melting and premelting. In addition to the conventional idea that bulk phase transitions are strongly affected by surfaces, we observed that surface behaviour is strongly affected by bulk transitions.

**Received 24 June 2015; accepted 8 January 2016.**
**Published online 14 March 2016.**

1. Dash, J. G., Rempel, A. W. & Wettlaufer, J. S. The physics of premelted ice and its geophysical consequences. *Rev. Mod. Phys.* **78,** 695–741 (2006).
2. Frenken, J. W. M. & van der Veen, J. F. Observation of surface melting. *Phys. Rev. Lett.* **54,** 134–137 (1985).
3. Zhu, D.-M., Pengra, D. & Dash, J. G. Edge melting in two-dimensional solid films. *Phys. Rev. B* **37,** 5586–5593 (1988).
4. Li, B., Zhou, D. & Han, Y. Assembly and phase transitions of colloidal crystals. *Nature Rev. Mater.* **1,** 15011 (2016).
5. Crocker, J. C. & Grier, D. G. Methods of digital video microscopy for colloidal studies. *J. Colloid Interface Sci.* **179,** 298–310 (1996).
6. Dash, J. G. History of the search for continuous melting. *Rev. Mod. Phys.* **71,** 1737–1743 (1999).
7. Chui, S. T. Grain-boundary theory of melting in two dimensions. *Phys. Rev. Lett.* **48,** 933–935 (1982).
8. Elbaum, M., Lipson, S. G. & Dash, J. G. Optical study of surface melting on ice. *J. Cryst. Growth* **129,** 491–505 (1993).
9. van der Gon, A. W. D., Gay, J. M., Frenken, J. W. M. & van der Veen, J. F. Order–disorder transitions at the Ge(111) surface. *Surf. Sci.* **241,** 335–345 (1991).
10. Carnevali, P., Ercolessi, F. & Tosatti, E. Melting and nonmelting behavior of the Au(111) surface. *Phys. Rev. B* **36,** 6701–6704 (1987).
11. Rogers, W. B. & Manoharan, V. N. Programming colloidal phase transitions with DNA strand displacement. *Science* **347,** 639–642 (2015).
12. Savage, J. R., Blair, D. W., Levine, A. J., Guyer, R. A. & Dinsmore, A. D. Imaging the sublimation dynamics of colloidal crystallites. *Science* **314,** 795–798 (2006).
13. Hertlein, C., Helden, L., Gambassi, A., Dietrich, S. & Bechinger, C. Direct measurement of critical Casimir forces. *Nature* **451,** 172–175 (2008).
14. Yang, Y., Asta, M. & Laird, B. B. Solid–liquid interfacial premelting. *Phys. Rev. Lett.* **110,** 096102 (2013).
15. Strandburg, K. J. Two-dimensional melting. *Rev. Mod. Phys.* **60,** 161–207 (1988).
16. Pluis, B., Frenkel, D. & van der Veen, J. F. Surface-induced melting and freezing II. A semi-empirical Landau-type model. *Surf. Sci.* **239,** 282–300 (1990).
17. Di Tolla, F. D. Interplay of melting, wetting, overheating and faceting on metal surfaces: theory and simulation. *Surf. Sci.* **377–379,** 499–503 (1997).
18. Karim, O. A. & Haymet, A. D. J. The ice/water interface: a molecular dynamics simulation study. *J. Chem. Phys.* **89,** 6889–6896 (1988).
19. Broughton, J. Q. & Gilmer, G. H. Interface melting: simulations of surfaces and grain boundaries at high temperatures. *J. Phys. Chem.* **91,** 6347–6359 (1987).
20. Lipowsky, R. Surface induced disordering at first-order bulk transitions. *Z. Phys. B* **51,** 165–172 (1983).
21. von Grünberg, H. H., Keim, P., Zahn, K. & Maret, G. Elastic behavior of a two-dimensional crystal near melting. *Phys. Rev. Lett.* **93,** 255703 (2004).
22. Still, T. *et al.* Phonon dispersion and elastic moduli of two-dimensional disordered colloidal packings of soft particles with frictional interactions. *Phys. Rev. E* **89,** 012301 (2014).
23. Lee, S. *et al.* Giant magneto-elastic coupling in multiferroic hexagonal manganites. *Nature* **451,** 805–808 (2008).
24. Bolhuis, P. & Frenkel, D. Prediction of an expanded-to-condensed transition in colloidal crystals. *Phys. Rev. Lett.* **72,** 2211–2214 (1994).
25. Alsayed, A. M., Islam, M. F., Zhang, J., Collings, P. J. & Yodh, A. G. Premelting at defects within bulk colloidal crystals. *Science* **309,** 1207–1210 (2005).
26. Halperin, B. I. & Nelson, D. R. Theory of two-dimensional melting. *Phys. Rev. Lett.* **41,** 121–124 (1978).
27. Zahn, K., Lenke, R. & Maret, G. Two-stage melting of paramagnetic colloidal crystals in two dimensions. *Phys. Rev. Lett.* **82,** 2721–2724 (1999).
28. Han, Y., Ha, N. Y., Alsayed, A. M. & Yodh, A. G. Melting of two-dimensional tunable-diameter colloidal crystals. *Phys. Rev. E* **77,** 041406 (2008).
29. Kapfer, S. C. & Krauth, W. Two-dimensional melting: from liquid-hexatic coexistence to continuous rransitions. *Phys. Rev. Lett.* **114,** 035702 (2015).
30. Nosenko, V., Zhdanov, S. K., Ivlev, A. V., Knapek, C. A. & Morfill, G. E. 2D melting of plasma crystals: equilibrium and nonequilibrium regimes. *Phys. Rev. Lett.* **103,** 015001 (2009).
31. Qi, W. & Dijkstra, M. Destabilisation of the hexatic phase in systems of hard disks by quenched disorder due to pinning on a lattice. *Soft Matter* **11,** 2852–2856 (2015).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank X. Cao and M. Liao for the discussions and Chromatech Inc. for the dye sample. This study was supported by RGC grants GRF601911, GRF16301514 and C6004-14G (Y.H.) and an NWO VENI grant 680-47-441 and the Start-Up Grant from Nanyang Technological University (R.N.).

**Author Contributions** Y.H. and B.L. conceived and designed the research. B.L. carried out the experiment with help from D.Z. and Y.P., and the data analysis with help from F.W. and R.N. B.L. and Y.H. wrote the paper. Y.H. supervised the work. All authors discussed the results.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Y.H. (yilong@ust.hk).

doi:10.1038/nature17151

# On-surface synthesis of graphene nanoribbons with zigzag edge topology

Pascal Ruffieux[1]*, Shiyong Wang[1]*, Bo Yang[2]*, Carlos Sánchez-Sánchez[1]*, Jia Liu[1]*, Thomas Dienel[1], Leopold Talirz[1], Prashant Shinde[1], Carlo A. Pignedoli[1,3], Daniele Passerone[1], Tim Dumslaff[2], Xinliang Feng[4], Klaus Müllen[2] & Roman Fasel[1,5]

**Graphene-based nanostructures exhibit electronic properties that are not present in extended graphene. For example, quantum confinement in carbon nanotubes and armchair graphene nanoribbons leads to the opening of substantial electronic bandgaps that are directly linked to their structural boundary conditions[1,2]. Nanostructures with zigzag edges are expected to host spin-polarized electronic edge states and can thus serve as key elements for graphene-based spintronics[3]. The edge states of zigzag graphene nanoribbons (ZGNRs) are predicted to couple ferromagnetically along the edge and antiferromagnetically between the edges[4], but direct observation of spin-polarized edge states for zigzag edge topologies—including ZGNRs—has not yet been achieved owing to the limited precision of current top-down approaches[5–10]. Here we describe the bottom-up synthesis of ZGNRs through surface-assisted polymerization and cyclodehydrogenation of specifically designed precursor monomers to yield atomically precise zigzag edges. Using scanning tunnelling spectroscopy we show the existence of edge-localized states with large energy splittings. We expect that the availability of ZGNRs will enable the characterization of their predicted spin-related properties, such as spin confinement[11] and filtering[12,13], and will ultimately add the spin degree of freedom to graphene-based circuitry.**

Graphene nanoribbons with armchair edges (AGNRs) can be synthesized using a bottom-up approach based on surface-assisted polymerization and subsequent cyclodehydrogenation of oligophenylene precursor monomers[14]. On-surface synthesis has been applied to fabricate many different AGNR structures[14–16], N-doped AGNRs[17,18] as well as AGNR heterostructures[18,19]. However, it is not directly suited for ZGNRs because polymerization of monomers via aryl–aryl coupling takes place along the armchair direction rather than the zigzag direction (Fig. 1a). Additionally, dehydrogenative cyclization of phenyl subgroups is not sufficient to form pure zigzag edges. Therefore, additional carbon functions must be placed at the edges of the monomers to complete the tiling toolbox needed for the bottom-up fabrication of arbitrary GNR structures.

In our protocol, surface-assisted polymerization and subsequent cyclization of suitably designed molecular precursors enables the synthesis of ZGNRs with control over both ribbon width and edge morphology. The idea depends on the choice of a U-shaped monomer (**1**), shown in Fig. 1b. With two halogen functions for thermally induced aryl–aryl coupling at the $R_1$ positions, this monomer allows surface-assisted polymerization to make a snake-like polymer. This monomer design can, in principle, be generalized to form wider structures by adding additional phenyl groups at the $R_1$ position. Furthermore, it affords accommodation of additional phenyl groups at the $R_2$ position that fill the holes in the interior of the undulating polymer. The crucial precursor is monomer **1a**, which carries two additional methyl groups. In this case, as well as the polymerization and planarization, an oxidative ring closure including the methyl groups is expected that would then establish two new six-membered rings together with the zigzag edge structure. The choice of $R_3$ further enables either the growth of pure zigzag edges ($R_3 = H$, **1a**) or the addition of further functional groups to the edges (**1b**).

Monomer **1a** was successfully obtained via multi-step organic synthesis (see Supplementary Information) and deposited on the clean Au(111) single-crystal surface by thermal sublimation under ultra-high vacuum conditions. If the surface is held at the dehalogenation temperature of 475 K, then precursor monomers are immediately activated and undergo polymer formation via radical addition (step 1 in Fig. 1c). Further annealing to the cyclodehydrogenation temperature of 625 K is then applied to form the final 6-ZGNR (that is, a ZGNR that is six carbon zigzag lines wide; step 2 in Fig. 1c). This two-step process has been successfully monitored by scanning tunnelling microscopy (STM; Fig. 2a, b). Large-scale STM images of the Au(111) surface after deposition of precursor monomer **1a** onto the substrate held at a temperature of 475 K reveal the formation of long (approximately 50 nm) polymers for which the meandering maxima in the apparent height have a periodicity of 1.55 nm, evidencing covalent-bond formation between the precursor monomers (Fig. 1c). The maxima with apparent heights of 0.3 nm are attributed to the sterically induced out-of-plane conformation of the phenyl ring carrying the methyl groups. Further annealing the sample to 625 K results in a complete planarization of the linear structures and a decrease in apparent height to 0.2 nm, consistent with the formation of the fully conjugated ribbon structure[14]. Small-scale images (inset of Fig. 2b) reveal completely smooth and flat edge areas. This indicates that, in addition to the cyclodehydrogenation of the two phenyl rings, the methyl groups are dehydrogenatively incorporated to form a fully conjugated system with atomically precise zigzag edges. Further structural details are accessible by non-contact atomic force microscopy (nc-AFM) imaging with a CO-functionalized tip (Fig. 2c), which enables direct imaging of the local bond configurations at small distances[20]. The achieved resolution directly confirms that the observed width and edge morphology correspond to the expected 6-ZGNR structure as defined by the design of **1a**. Furthermore, we can unambiguously state that the zigzag-edge atoms have the expected mono-hydrogen termination. Other possible terminations such as radical edges due to complete dehydrogenation or $H_2$ termination can be disregarded, owing to the absence of bending across the ribbon (which relates to bonding of the radical edges to the substrate[21]) and of the distinct maxima related to $H_2$ edge termination (see Supplementary Fig. 2), respectively. Thus, our 6-ZGNRs exhibit atomically precise edges with the expected CH termination. Both features are crucial to host the predicted antiferromagnetic edge states.

As can be seen from the STM image in Fig. 2b, the chemistry of the ZGNR fabrication process faces intrinsic complications such as frequent thermally induced chemical cross-linking of ribbons during

[1]Empa, Swiss Federal Laboratories for Materials Science and Technology, 8600 Dübendorf, Switzerland. [2]Max Planck Institute for Polymer Research, 55128 Mainz, Germany. [3]NCCR MARVEL, Empa, Swiss Federal Laboratories for Materials Science and Technology, 8600 Dübendorf, Switzerland. [4]Center for Advancing Electronics Dresden & Department of Chemistry and Food Chemistry, Technische Universität Dresden, 01062 Dresden, Germany. [5]Department of Chemistry and Biochemistry, University of Bern, 3012 Bern, Switzerland.
*These authors contributed equally to this work.

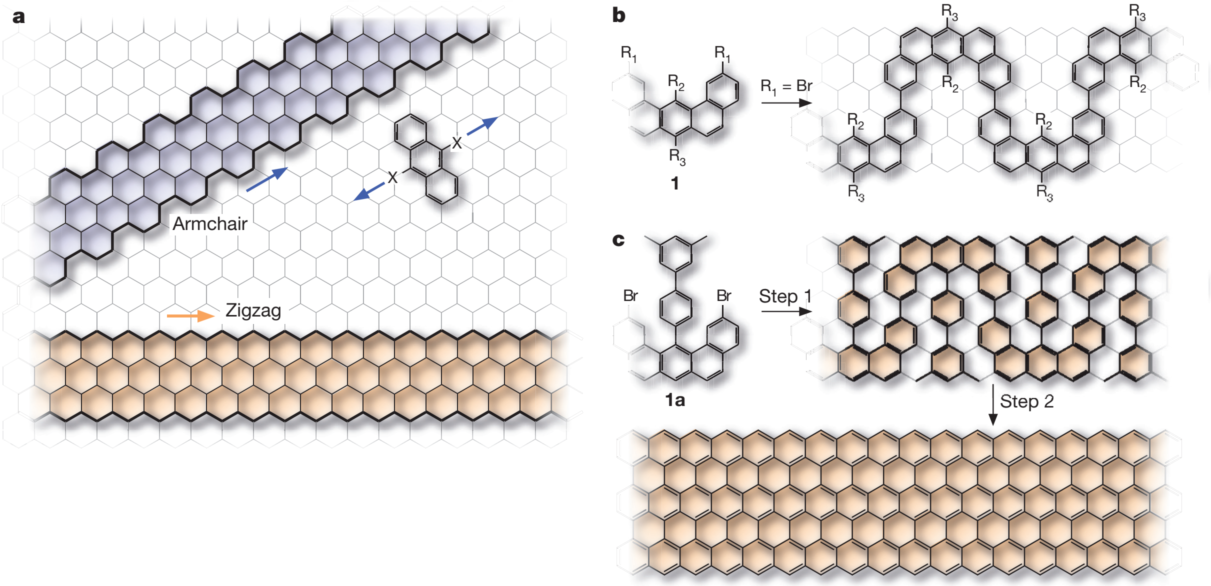


**Figure 1 | Synthetic strategy to GNRs with zigzag edges. a**, Structure of armchair and zigzag graphene nanoribbons, and an exemplary anthracene-based molecular precursor for the bottom-up fabrication of AGNRs via aryl–aryl coupling (X = halogen). **b**, U-shaped dibenzo[a,j] anthracene monomer **1** with halogen functions $R_1 = Br$ designed to enable surface-assisted aryl–aryl coupling into a snake-like polymer along the zigzag direction. **c**, Monomer **1a**, with an additional dimethyl-biphenyl group in the interior of the U-shape ($R_2$ position), which is designed to afford a 6-ZGNR upon polymerization (step 1) and subsequent cyclization (step 2).

the cyclization step, and, more severely, strong electronic coupling between the ribbons and the metal surface that obscures the detection of the electronic edge states. In fact, no evidence for increased intensity at the zigzag edges could be obtained in differential-conductance (d*I*/d*V*) maps recorded with tunnelling resistances of more than 0.6 MΩ, and spectra taken above the 6-ZGNRs are dominated by the (up-shifted) surface state of the underlying Au(111) substrate (not shown). However, unambiguous evidence for the sought-after edge states is obtained for 6-ZGNRs manipulated with the STM tip onto post-deposited insulating NaCl islands, where they are electronically decoupled from the underlying metal substrate[22]. An example STM image of a 6-ZGNR bridging between two NaCl islands is shown in Fig. 3a; a d*I*/d*V* spectrum taken at the edge of the decoupled ZGNR segment is shown in Fig. 3b. In contrast to the result on Au(111), this spectrum clearly exhibits three resonance peaks near the Fermi level, with energy splittings of $\Delta^0 = 1.5$ eV and $\Delta^1 = 1.9$ eV between the two occupied states and the unoccupied one (Fig. 3b). d*I*/d*V* maps acquired at these peaks demonstrate that the corresponding states are highly localized at the zigzag edges (Fig. 3d). Their characteristic features, such as a protrusion at each outermost zigzag carbon atom and an enhanced intensity at the ribbon terminus, are in excellent agreement with the local density of states (DOS) of the corresponding Kohn–Sham density functional theory (DFT) orbitals shown in Fig. 3e. Although effective mean-field theories, such as Kohn–Sham DFT, tend to provide reliable information about the energy level ordering and the shape of orbitals in graphene nanostructures, the same is not true for the size of the electronic gap. More accurate predictions of the bandgaps of graphene nanoribbons are obtained using the *GW* approximation of many-body perturbation theory[23] (where '*G*' and '*W*' stand for 'Green's function' and 'screened interaction', respectively).

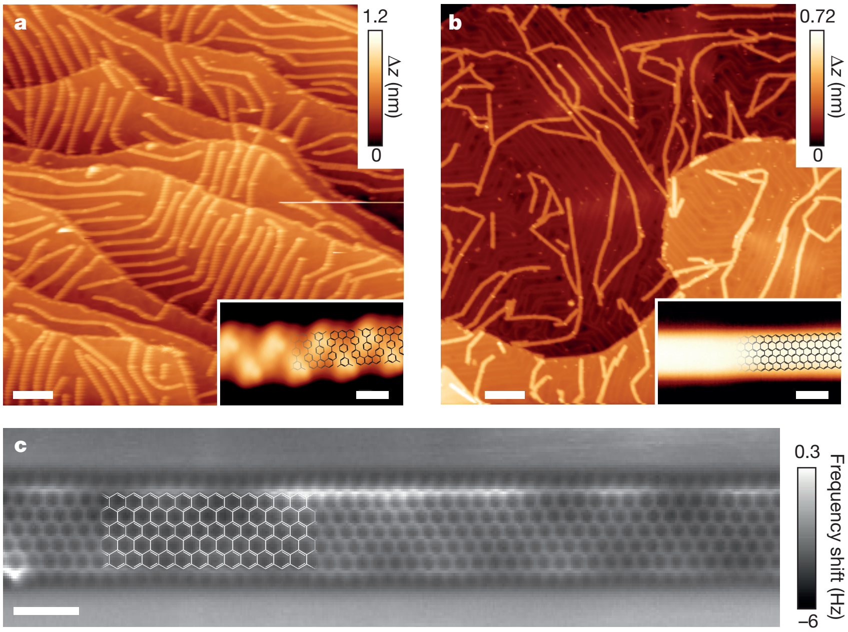


**Figure 2 | Synthesis and characterization of atomically precise 6-ZGNRs. a**, Large-scale STM image of the Au(111) surface after deposition of monomer **1a** on the surface held at 475 K. Formation of snake-like polymers is observed. Scanning parameters: voltage $V = -1.5$ V, current $I = 40$ pA, colour-scale range $\Delta z = 1.2$ nm. Scale bar, 20 nm. Inset, high-resolution STM image of the polymer. Zigzag alternation of bright maxima indicates the lifting and/or tilting of the phenyl rings carrying the methyl groups. A structural model is superimposed for comparison. $V = -1.3$ V, $I = 10$ pA, $\Delta z = 0.3$ nm. Scale bar, 1 nm. **b**, Large-scale STM image of the Au(111) surface after annealing at 625 K. The flatter appearance, reduced apparent height and lack of internal structure indicate the complete cyclodehydrogenation of the polymers and the formation of 6-ZGNRs. $V = -1.0$ V, $I = 20$ pA, $\Delta z = 0.7$ nm. Scale bar, 20 nm. Inset, high-resolution STM of a 6-ZGNR, which is in excellent agreement with the superimposed structural model. $V = -0.3$ V, $I = 5$ pA, $\Delta z = 0.2$ nm. Scale bar, 1 nm. **c**, Constant height nc-AFM frequency shift image taken with a CO-functionalized tip. The intra-ribbon resolution shows the formation of a 6-ZGNR with atomically precise CH edges. A $CH_2$ defect is seen in the lower left corner. Oscillation amplitude $A_{osc} = 0.7$ Å, $V = 5$ mV. Scale bar, 1 nm.

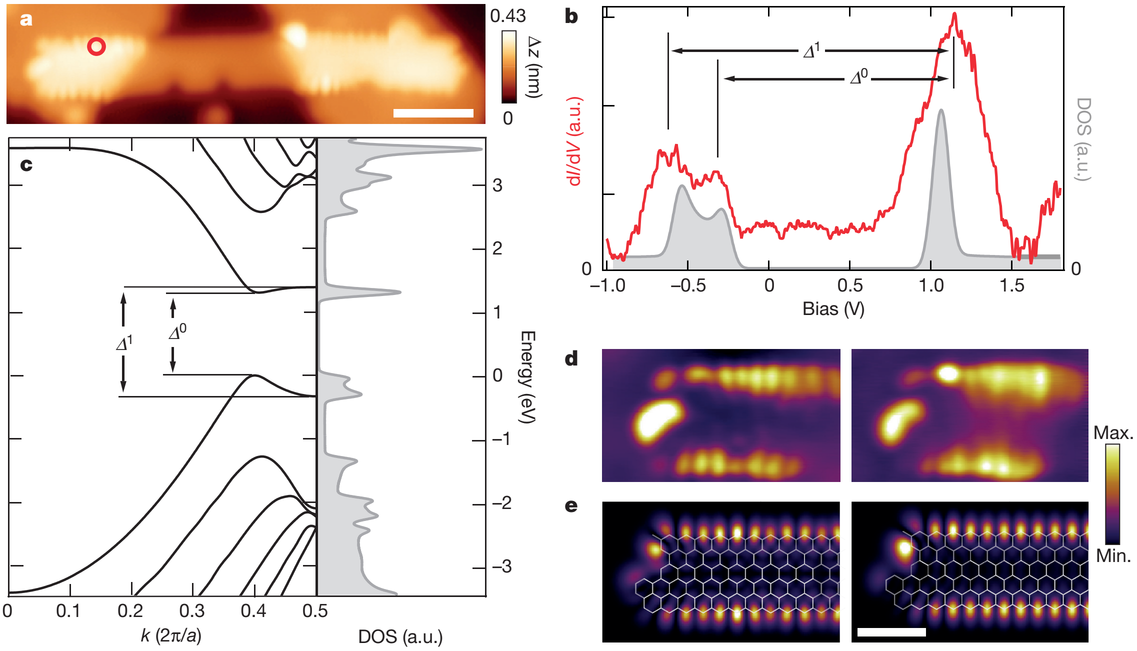


**Figure 3 | Edge-state characterization of 6-ZGNRs. a**, STM topography image ($V = -0.25$ V, $I = 100$ pA) of a 6-ZGNR bridging between two NaCl monolayer islands, achieved through STM manipulation. Scale bar, 2 nm. **b**, Differential conductance ($dI/dV$) spectrum (red) taken at the zigzag edge marked by the red circle in **a** and the quasiparticle density of states (DOS; grey). **c**, Quasiparticle band structure (energy versus wave vector k) (left, black; *a* is the lattice parameter) and DOS (right, grey) calculated for an infinitely long 6-ZGNR. **d**, Differential-conductance maps of filled (left) and empty (right) edge states taken at a sample bias of −0.3 V and 1.0 V, respectively. **e**, DFT-based local DOS at a 4-Å tip–sample distance, showing the spatial distribution of filled (left, with overlaid structural model) and empty (right, with overlaid structural model) edge states. Scale bar for **d** and **e**, 1 nm. a.u., arbitrary units.

The resultant quasiparticle band structure and the corresponding DOS are presented in Fig. 3c. The calculated energy splittings ($\Delta^0 = 1.4$ eV and $\Delta^1 = 1.7$ eV) are in good agreement with experiment. (The quantitative agreement between the experimentally obtained energy splittings and those calculated using many-body perturbation theory is attributed to a compensation of bandgap reduction due to screening by the underlying metal and NaCl island, and to bandgap increase due to quantum confinement within the finite decoupled ribbon segments.) This agreement, in conjunction with the agreement between experimentally determined and simulated DOS maps, clearly evidences the experimental observation of the unperturbed zigzag edge states. Although edge states have previously been observed in several systems with less-well-defined zigzag edges[9,24–27], the reported energy splittings vary greatly and are substantially smaller than those reported

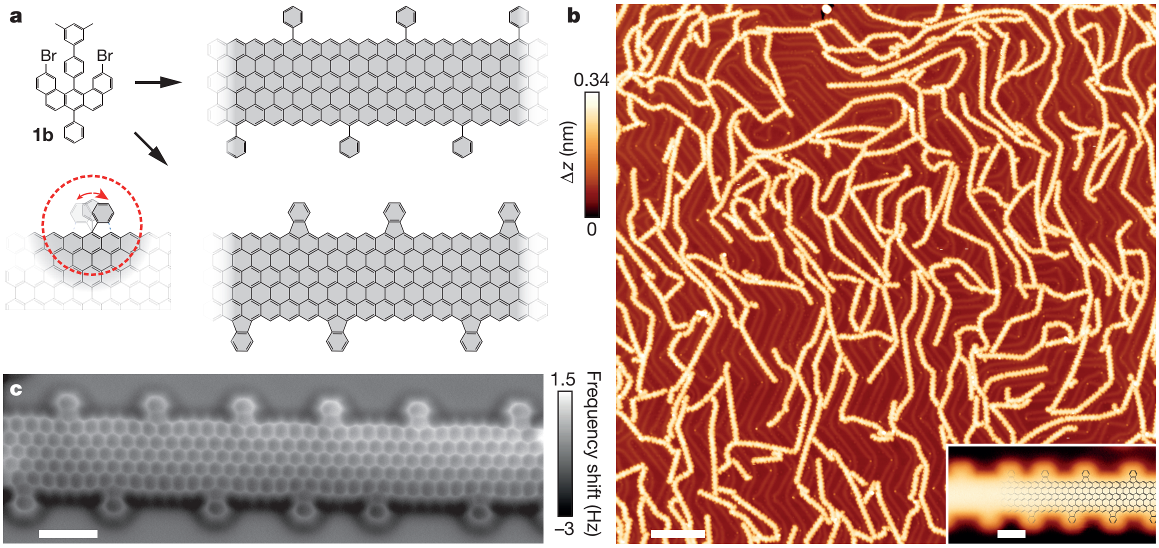


**Figure 4 | Synthesis and characterization of edge-modified 6-ZGNRs. a**, Monomer **1b** (top left) with an additional phenyl group at the $R_3$ position of the monomer **1a**, which is designed to afford an edge-modified 6-ZGNR upon polymerization and subsequent cyclization. The highlighted scheme (dashed red circle) illustrates the two possible rotation possibilities (bottom left, red arrow) upon activation of the external phenyl ring. Right, possible cyclodehydrogenation products assuming no activation of the external phenyl groups (top) and formation of fluoranthene subunits based on an additional dehydrogenative ring closure at the external phenyl groups (bottom). **b**, Overview STM image ($V = -1.5$ V, $I = 150$ pA) of edge-modified 6-ZGNRs fabricated on a Au(111) surface. Scale bar, 20 nm. Inset, high-resolution STM image ($V = 0.15$ V, $I = 2$ pA). Scale bar, 1 nm. **c**, Constant-height nc-AFM frequency-shift image of edge-modified 6-ZGNR ($A_{osc} = 0.7$ Å, $V = 25$ mV). Scale bar, 1 nm.

here. This discrepancy indicates that the electronic structure of zigzag edges is extremely sensitive to edge roughness and interaction with the supporting substrate.

The synthesis of GNRs with perfect zigzag edge periphery combined with convincing evidence of their edge states could lead to new methods for experimentally verifying the electronic, optical and magnetic properties predicted by theory[28–30], and to the systematic engineering of these properties by using modified types of ZGNRs. As a first step along these lines, we further refined our monomer design by introducing an analogous compound **1b** (Fig. 4), which is similar to **1a** except that it bears an additional phenyl group at the $R_3$ position (see Supplementary Information). We expected that, owing to the steric hindrance brought about by the twisted phenyl group, the growing ribbons would be more efficiently decoupled from the surface and potentially better shielded from neighbouring GNRs. The structural characterization of the ZGNRs obtained from monomer **1b** in an analogous, thermally induced polymerization–cyclization procedure is shown in Fig. 4b, c. The nc-AFM images clearly reveal that, at the cyclodehydrogenation temperature of 573 K, the external phenyl group undergoes a ring closure under the formation of a fluoranthene-type subunit with an incorporated five-membered ring, as illustrated in Fig. 4a. Because the additional ring closure can occur by dehydrogenation of either of the neighbouring zigzag-edge carbon atoms, we do not expect to observe fully periodic arrangement of the fluoranthene subunits. This expectation is confirmed by the nc-AFM image shown in Fig. 4c, which reveals three, four and five zigzag cusps that separate neighbouring fluoranthene subunits.

Controlled edge modification is a useful strategy for engineering the band structure of edge states. However, the emphasis here is on the reduction of ZGNR–substrate interaction. Indeed, the edge modification discussed above alters the ZGNR–substrate interaction enough to enable the STM to map the typical features of the edge state (see Supplementary Fig. 7).

The successful bottom-up synthesis of atomically precise ZGNRs opens extensive opportunities and challenges for the analysis of their physical properties (for example, their band structure, magnetism and charge/spin transport) and for the fabrication of ZGNR-based devices such as the proposed spin valves[12]. However, the strong interaction of the pristine 6-ZGNR with the metal substrate (as reflected in the obstruction of the spectral features of the edge states) raises important questions regarding the chemical reactivity of the zigzag edges, which needs to be controlled to be able to study and apply these materials under ambient conditions. A promising method for doing so is provided by edge 'passivation' approaches that preserve the electronic properties of the ribbon, but reduce the chemical reactivity of their (functionalized) edges[31].

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 7 July 2015; accepted 19 January 2016.**

1. Saito, R., Fujita, M., Dresselhaus, G. & Dresselhaus, M. S. Electronic structure of graphene tubules based on $C_{60}$. *Phys. Rev. B* **46,** 1804–1811 (1992).
2. Wakabayashi, K., Fujita, M., Ajiki, H. & Sigrist, M. Electronic and magnetic properties of nanographite ribbons. *Phys. Rev. B* **59,** 8271–8282 (1999).
3. Han, W., Kawakami, R. K., Gmitra, M. & Fabian, J. Graphene spintronics. *Nature Nanotechnol.* **9,** 794–807 (2014).
4. Nakada, K., Fujita, M., Dresselhaus, G. & Dresselhaus, M. S. Edge state in graphene ribbons: nanometer size effect and edge shape dependence. *Phys. Rev. B* **54,** 17954–17961 (1996).
5. Han, M. Y., Ozyilmaz, B., Zhang, Y. & Kim, P. Energy band-gap engineering of graphene nanoribbons. *Phys. Rev. Lett.* **98,** 206805 (2007).
6. Kosynkin, D. V. *et al.* Longitudinal unzipping of carbon nanotubes to form graphene nanoribbons. *Nature* **458,** 872–876 (2009).
7. Li, X., Wang, X., Zhang, L., Lee, S. & Dai, H. Chemically derived, ultrasmooth graphene nanoribbon semiconductors. *Science* **319,** 1229–1232 (2008).
8. Wang, X. & Dai, H. Etching and narrowing of graphene from the edges. *Nature Chem.* **2,** 661–665 (2010).
9. Magda, G. Z. *et al.* Room-temperature magnetic order on zigzag edges of narrow graphene nanoribbons. *Nature* **514,** 608–611 (2014).
10. Ma, L., Wang, J. & Ding, F. Recent progress and challenges in graphene nanoribbon synthesis. *ChemPhysChem* **14,** 47–54 (2013).
11. Topsakal, M., Sevinçli, H. & Ciraci, S. Spin confinement in the superlattices of graphene ribbons. *Appl. Phys. Lett.* **92,** 173118 (2008).
12. Wimmer, M., Adagideli, İ., Berber, S., Tománek, D. & Richter, K. Spin currents in rough graphene nanoribbons: universal fluctuations and spin injection. *Phys. Rev. Lett.* **100,** 177207 (2008).
13. Son, Y.-W., Cohen, M. L. & Louie, S. G. Half-metallic graphene nanoribbons. *Nature* **444,** 347–349 (2006).
14. Cai, J. *et al.* Atomically precise bottom-up fabrication of graphene nanoribbons. *Nature* **466,** 470–473 (2010).
15. Chen, Y.-C. *et al.* Tuning the band gap of graphene nanoribbons synthesized from molecular precursors. *ACS Nano* **7,** 6123–6128 (2013).
16. Zhang, H. *et al.* On-surface synthesis of rylene-type graphene nanoribbons. *J. Am. Chem. Soc.* **137,** 4022–4025 (2015).
17. Bronner, C. *et al.* Aligning the band gap of graphene nanoribbons by monomer doping. *Angew. Chem. Int. Ed.* **52,** 4422–4425 (2013).
18. Cai, J. *et al.* Graphene nanoribbon heterojunctions. *Nature Nanotechnol.* **9,** 896–900 (2014).
19. Chen, Y.-C. *et al.* Molecular bandgap engineering of bottom-up synthesized graphene nanoribbon heterojunctions. *Nature Nanotechnol.* **10,** 156–160 (2015).
20. Gross, L., Mohn, F., Moll, N., Liljeroth, P. & Meyer, G. The chemical structure of a molecule resolved by atomic force microscopy. *Science* **325,** 1110–1114 (2009).
21. Li, Y., Zhang, W., Morgenstern, M. & Mazzarello, R. Electronic and magnetic properties of zigzag graphene nanoribbons on the (111) surface of Cu, Ag, and Au. *Phys. Rev. Lett.* **110,** 216804 (2013).
22. Repp, J., Meyer, G., Stojković, S. M., Gourdon, A. & Joachim, C. Molecules on insulating films: scanning-tunneling microscopy imaging of individual molecular orbitals. *Phys. Rev. Lett.* **94,** 026803 (2005).
23. Yang, L., Park, C.-H., Son, Y.-W., Cohen, M. L. & Louie, S. G. Quasiparticle energies and band gaps in graphene nanoribbons. *Phys. Rev. Lett.* **99,** 186801 (2007).
24. Ritter, K. A. & Lyding, J. W. The influence of edge structure on the electronic properties of graphene quantum dots and nanoribbons. *Nature Mater.* **8,** 235–242 (2009).
25. Tao, C. *et al.* Spatially resolving edge states of chiral graphene nanoribbons. *Nature Phys.* **7,** 616–620 (2011).
26. van der Lit, J. *et al.* Suppression of electron–vibron coupling in graphene nanoribbons contacted via a single atom. *Nature Commun.* **4,** 2023 (2013).
27. Li, Y. Y., Chen, M. X., Weinert, M. & Li, L. Direct experimental determination of onset of electron–electron interactions in gap opening of zigzag graphene nanoribbons. *Nature Commun.* **5,** 4311 (2014).
28. Dutta, S. & Wakabayashi, K. Tuning charge and spin excitations in zigzag edge nanographene ribbons. *Sci. Rep.* **2,** 519 (2012).
29. Yang, L., Cohen, M. L. & Louie, S. G. Magnetic edge-state excitons in zigzag graphene nanoribbons. *Phys. Rev. Lett.* **101,** 186401 (2008).
30. Yazyev, O. V. A guide to the design of electronic properties of graphene nanoribbons. *Acc. Chem. Res.* **46,** 2319–2328 (2013).
31. Li, Y., Zhou, Z., Cabrera, C. R. & Chen, Z. Preserving the edge magnetism of zigzag graphene nanoribbons by ethylene termination: insight by Clar's rule. *Sci. Rep.* **3,** 2030 (2013).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** This work was supported by the Swiss National Science Foundation, the Office of Naval Research BRC Program, the European Research Council (grant NANOGRAPH), the DFG Priority Program SPP 1459, the Graphene Flagship (No. CNECT-ICT-604391), and the European Union Projects UPGRADE, GENIUS and MoQuaS. We acknowledge the Swiss Supercomputing Center (CSCS) for computational resources (project s507). We thank A. Ferretti for his contribution to this project and O. Gröning for discussions.

**Author Contributions** P.R., R.F., X.F. and K.M. conceived and supervised the experiments. B.Y. and T.Du. synthesized the precursor monomers. J.L. and C.S. developed the on-surface synthesis protocols and did the STM analysis. T.Di., J.L. and S.W. performed the AFM imaging; S.W. and J.L. did the spectroscopic analysis. P.S., L.T., C.A.P. and D.P. performed the simulations. C.S., S.W. and P.R. made the figures. P.R., K.M. and R.F. wrote the paper. All authors discussed the results and implications and commented on the manuscript at all stages.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.F. (roman.fasel@empa.ch) or K.M. (muellen@mpip-mainz.mpg.de).

## METHODS

**Sample preparation.** Experiments were carried out under ultra-high vacuum conditions (base pressure $10^{-11}$ mbar) with a low-temperature STM and a low-temperature STM/AFM, both from Omicron-Oxford. A Au(111) single crystal was used as the substrate for the growth of 6-ZGNRs and edge-modified 6-ZGNRs. The Au(111) surface was cleaned by repeated cycles of argon-ion bombardment and annealing at 750 K for 15 min until a clean surface was obtained, as judged by STM. Molecular precursors were thermally deposited on the clean Au(111) surface held at room temperature with a typical rate of 1 Å $\text{min}^{-1}$. After deposition, the sample was post-annealed in four steps to 475 K, 525 K, 575 K and 625 K for typically 15 min per step to achieve long, high-quality GNRs. A modified growth protocol was used for the experiments in which 6-ZGNRs are transferred onto NaCl islands. After precursor deposition at room temperature, the substrate temperature was directly increased to 625 K (1 K $\text{s}^{-1}$) and kept at this temperature for 5 min. This protocol increases the number of short and isolated 6-ZGNRs needed for successful manipulation with the STM tip. Compared to the protocol described above, this increases the number of defects at the ZGNR edges.

**Imaging, manipulation and spectroscopy.**

*Constant-current STM imaging.* STM images were acquired in the constant-current mode at sample temperatures of 77 K or 5 K, as indicated in each case. Scanning parameters are specified in each figure caption.

*Constant-height, non-contact AFM imaging.* Non-contact AFM measurements were performed with a tungsten tip attached to a tuning fork sensor[32]. The tip was *a posteriori* functionalized by the controlled adsorption of a single CO molecule at the tip apex from the previously CO-dosed surface[33]. This procedure enables the imaging of the chemical structure of organic molecules[20]. The sensor was driven close to its resonance frequency (about 23,570 Hz) with a constant amplitude of approximately 70 pm. The shift in the resonance frequency of the tuning fork (with the attached CO-functionalized tip) was recorded in constant-height mode (Omicron Matrix electronics and HF2Li PLL by Zurich Instruments).

*Transfer of GNRs onto NaCl monolayer islands.* We developed a routine to transfer bottom-up fabricated short GNRs from the metal surface onto insulating NaCl islands in order to be able to access their intrinsic electronic structure. This method, in which physisorbed individual 6-ZGNRs are laterally and/or vertically manipulated on the Au(111) surface, consists of four steps: (1) deposition of NaCl islands (thermal evaporation of submonolayer coverage, deposition temperature of about 1,000 K) on the 6-ZGNR/Au(111) surface; (2) pick-up of one end of a GNR by approaching and retracting the STM tip with low bias (about −50 mV); (3) lateral displacement of the tip, together with the GNR, above the NaCl island; and (4) release of the ribbon by a 3.0 V voltage pulse, leaving the GNR partially adsorbed on NaCl and partially on the metal surface.

*Differential-conductance spectroscopy.* The differential conductance ($\text{d}I/\text{d}V$) measurements were performed in a low-temperature STM at 5 K via the lock-in technique, using a bias-voltage modulation of 20 mV and a frequency of 860 Hz. $\text{d}I/\text{d}V$ maps were acquired in the constant-current mode for the decoupled 6-ZGNR, and in the constant-height mode for ribbons on Au(111).

**Theoretical details.** DFT calculations were carried out using the CP2K code (http://www.cp2k.org) for geometry optimizations; the Quantum ESPRESSO code[34] was used for scanning tunnelling spectroscopy simulations. In CP2K, the core electrons and nuclei are represented using the GTH pseudopotential[35] and the valence electrons are treated with a triple-ζ valence basis set with two sets of $p$-type or $d$-type polarization functions (TZV2P)[36]; in Quantum ESPRESSO, a plane-wave basis set and norm-conserving pseudopotentials are used. The exchange-correlation was approximated by the PBE functional[37].

The edge-modified 6-ZGNR structures shown in Fig. 4 were set up starting from the optimized atomic structure of the 6-ZGNR (lattice parameter $a = 2.461$ Å) and contain 24 unit cells. A 19-Å-wide (15-Å-wide) region of vacuum was included along the transverse (perpendicular) directions to avoid interactions between periodic replicas. Forces on the nuclei were then relaxed until all forces dropped below 5 meV $\text{Å}^{-1}$. Using the optimized atomic structure, a self-consistent field calculation was carried out in Quantum ESPRESSO, using energy cut-offs of 120 Ry and 480 Ry (1 Ry ≈ 13.6 eV) for the wave function and the charge density, respectively, and a $k$-point grid of $3 \times 1 \times 1$, including the Γ-point.

Quasi-particle corrections were computed within the framework of many-body perturbation theory, using a single iteration ($G_0W_0$) in the $GW$ approximation to the self-energy as implemented in the yambo code[38]. The electronic structure of the 6-ZGNRs from DFT was recalculated using a 60-Ry plane-wave cut-off, 64 $k$-points in the first Brillouin zone and 250 bands, covering the energy range up to 21 eV above the highest occupied band. The dielectric matrix ($\varepsilon$) was calculated in the random phase approximation with an 8-Ry cut-off for the plane-wave basis. $\varepsilon^{-1}$ was evaluated at frequencies of $\omega = 0$ and $\omega = i2$ Ry and extended to the real frequency axis using the plasmon-pole model by ref. 39. A rectangular Coulomb cut-off was used along the directions perpendicular to the GNR axis, as described in ref. 40.

32. Giessibl, F. J. Atomic resolution on Si(111)-(7 × 7) by noncontact atomic force microscopy with a force sensor based on a quartz tuning fork. *Appl. Phys. Lett.* **76,** 1470–1472 (2000).
33. Bartels, L. *et al.* Dynamics of electron-induced manipulation of individual CO molecules on Cu(111). *Phys. Rev. Lett.* **80,** 2004–2007 (1998).
34. Giannozzi, P. *et al.* QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21,** 395502 (2009).
35. Goedecker, S., Teter, M. & Hutter, J. Separable dual-space Gaussian pseudopotentials. *Phys. Rev. B* **54,** 1703–1710 (1996).
36. VandeVondele, J. & Hutter, J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *J. Chem. Phys.* **127,** 114105 (2007).
37. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77,** 3865 (1996).
38. Marini, A., Hogan, C., Grüning, M. & Varsano, D. yambo: an *ab initio* tool for excited state calculations. *Comput. Phys. Commun.* **180,** 1392–1403 (2009).
39. Godby, R. W. & Needs, R. J. Metal-insulator transition in Kohn-Sham theory and quasiparticle theory. *Phys. Rev. Lett.* **62,** 1169–1172 (1989).
40. Rozzi, C. A., Varsano, D., Marini, A., Gross, E. K. U. & Rubio, A. Exact Coulomb cutoff technique for supercell calculations. *Phys. Rev. B* **73,** 205119 (2006).

doi:10.1038/nature17149

# The past, present and future of African dust

Amato T. Evan[1,2], Cyrille Flamant[2], Marco Gaetani[2] & Françoise Guichard[3]

**African dust emission and transport exhibits variability on diurnal[1] to decadal[2] timescales and is known to influence processes such as Amazon productivity[3], Atlantic climate modes[4], regional atmospheric composition and radiative balance[5] and precipitation in the Sahel[6]. To elucidate the role of African dust in the climate system, it is necessary to understand the factors governing its emission and transport. However, African dust is correlated with seemingly disparate atmospheric phenomena, including the El Niño/Southern Oscillation[7,8], the North Atlantic Oscillation[9], the meridional position of the intertropical convergence zone[10,11], Sahelian rainfall[8] and surface temperatures over the Sahara Desert[12], all of which obfuscate the connection between dust and climate. Here we show that the surface wind field responsible for most of the variability in North African dust emission reflects the topography of the Sahara, owing to orographic acceleration of the surface flow. As such, the correlations between dust and various climate phenomena probably arise from the projection of the winds associated with these phenomena onto an orographically controlled pattern of wind variability. A 161-year time series of dust from 1851 to 2011, created by projecting this wind field pattern onto surface winds from a historical reanalysis[13], suggests that the highest concentrations of dust occurred from the 1910s to the 1940s and the 1970s to the 1980s, and that there have been three periods of persistent anomalously low dust concentrations—in the 1860s, 1950s and 2000s. Projections of the wind pattern onto climate models give a statistically significant downward trend in African dust emission and transport as greenhouse gas concentrations increase over the twenty-first century, potentially associated with a slow-down of the tropical circulation. Such a dust feedback, which is not represented in climate models, may be of benefit to human and ecosystem health in West Africa via improved air quality[14] and increased rainfall[6]. This feedback may also enhance warming of the tropical North Atlantic[15], which would make the basin more suitable for hurricane formation and growth[16].**

We perform an eigenanalysis of zonal and meridional wind speeds at 10 m above the surface from the European Centre for Medium-Range Weather Forecasts Interim reanalysis product[17] (ERA-I) to identify coherent variability in wind fields associated with dust emission and transport (see Methods). Most relevant to this study is the second empirical orthogonal function (EOF) and corresponding principal component (PC) time series from the eigenanalysis of 10-m winds (Fig. 1a), which explains approximately 20% of the variance in the data (EOFs and PCs for the first and third modes are shown in Extended Data Fig. 1). The spatial structure of the second EOF maximizes in the region of 15° N and 10°–20° E, with secondary maxima extending towards the northwest (Fig. 1a). The wind fields in this EOF exhibit a northeasterly flow across much of the Sahara, characteristic of the trade winds, as well as a westerly flow near 30° N and 0° E.

The corresponding second PC time series (PC2) has a maximum in the 1980s, which is followed by a steep decline over the following decade (Fig. 1b). From 2000 through to the end of the record, values for the ERA-I PC2 largely remain between 0 and −1 standard deviations. The time series of ERA-I PC2 is strikingly similar to that of dust optical depth over the Cape Verde islands (15° N, 23.5° W) retrieved from the Advanced Very High Resolution Radiometer (AVHRR) space-borne imager[2]. We average over this area in order to compare to coral proxy data[2], but note that dust over Cape Verde is highly representative of dust over the entire tropical North Atlantic[2] and our results are qualitatively identical if we instead average the AVHRR data over the entire tropical North Atlantic. In Fig. 1b both monthly time series are smoothed with a 13-month running mean filter to highlight variability on interannual and longer time scales. The correlation between the ERA-I PC2 and AVHRR dust smoothed time series is 0.76 ($P \approx 0.01$; all $P$ values reported here account for time series autocorrelation). The ERA-I wind fields and dust retrievals from the AVHRR are independent, so the fact that ERA-I PC2 explains 58% of the variance in the AVHRR dust data is strong evidence that dust emission and transport, on these temporal and spatial scales, can also be approximated as a linear function of surface wind speeds over the Sahara (see Methods and Extended Data Fig. 2).

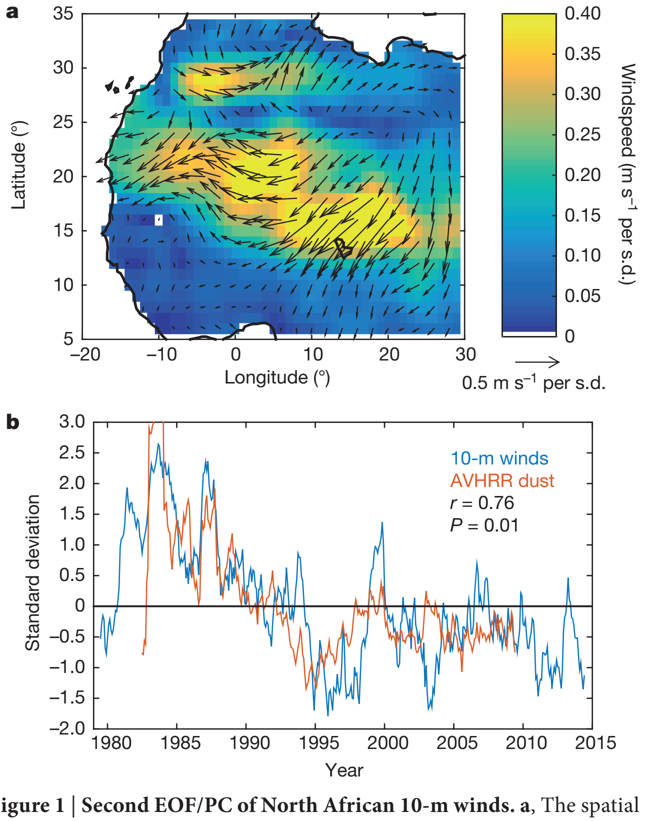


**Figure 1 | Second EOF/PC of North African 10-m winds. a**, The spatial structure of the second EOF of 10-m winds over North Africa. Arrows and shading represent the direction and magnitude of monthly mean winds, respectively, in units of wind speed per unit standard deviation (s.d.) change in the PC time series. **b**, The associated PC time series (10-m winds) and a time series of dust optical depth averaged over the tropical North Atlantic (AVHRR dust). The second EOF/PC explains 15% of the total variance in the surface winds data.

[1]Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California, USA. [2]Laboratoire Atmosphères, Milieux, Observations Spatiales (LATMOS)/IPSL, UPMC Université Paris 06, Sorbonne Université, UVSQ, CNRS, Paris, France. [3]CNRM-GAME, UMR 3589 CNRS and Météo-France, Toulouse, France.

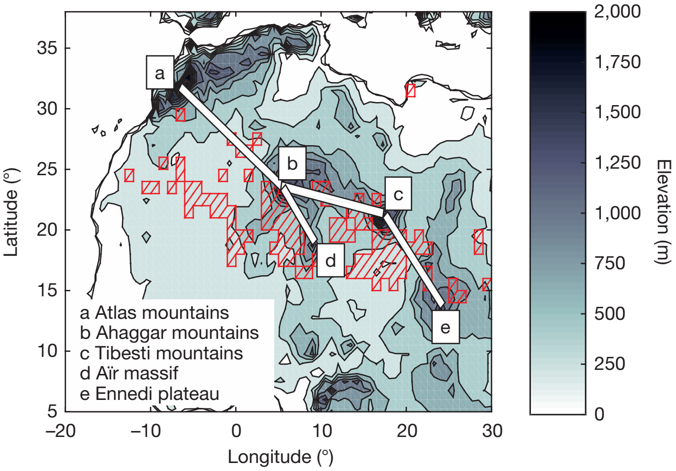


**Figure 2 | North African orography and dust source regions.** Shading represents surface elevation and transects indicate regions where orography increases the magnitude of the surface wind speed; the relevant topographic features are indicated in the legend and the surface altitude along each transect is shown in Extended Data Fig. 3. The red-hatched regions encompass the major Saharan dust sources[20].

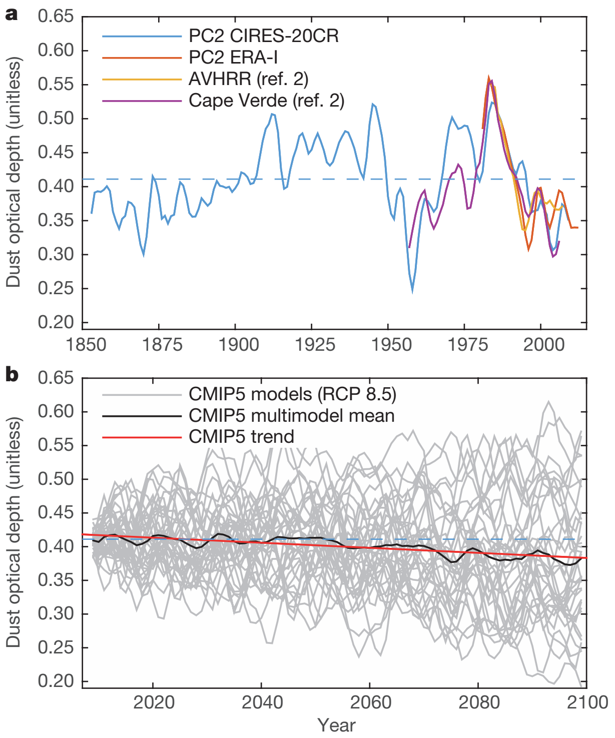


**Figure 3 | Estimates of North Atlantic dust. a**, PC2 time series from the CIRES-20CR and ERA-I reanalyses, and the AVHRR and hybrid satellite-coral (Cape Verde) dust time series. The ERA-I PC2 data are scaled to be in units of dust optical depth in a manner identical to that for the CIRES-20CR PC2 time series. **b**, The CMIP5 ensemble mean PC2 time series, the multimodel mean CMIP5 PC2 time series and its linear trend. The dashed blue lines are the long-term mean of the CIRES-20CR PC2 time series. Annual mean time series are smoothed with a 1-4-6-4-1 filter.

The spatial structure of the second EOF (Fig. 1a) closely resembles the topography over the Sahara (Fig. 2). In particular, the wind vectors exhibit a maximum in magnitude downstream of the gaps between the major Saharan mountains and plateaus, which are indicated by the transect lines and labels a–e (surface elevations are shown for each transect in Extended Data Fig. 3). It is not surprising that this EOF of 10-m winds reflects surface topography; flow acceleration at the exit region of a gap results in an increase of the variance of the down-gap winds, relative to variance of the up-gap, non-accelerated winds[18]. This similarity in the spatial structure of the second EOF and the orography of the Sahara, and the high correlation between the PC2 and the AVHRR dust time series, are both consistent with previous work showing that most dust emitted from the Sahara is generated within major topographic depressions[19]. According to a recent study[20], approximately 85% of all North African dust emission occurs within the areas indicated with blue hatching, which are generally downwind of the gaps indicated in Fig. 2, and which encompasses the regions where the magnitude of the second EOF is high.

We next project the spatial structure of the second EOF (Fig. 1a) onto the NOAA-CIRES 20th Century Reanalysis[21] (CIRES-20CR) monthly mean 10-m wind fields to recreate a historical proxy record of dust emission and westward transport over the Atlantic. We convert the units of the CIRES-20CR PC2 time series to dust optical depth by linearly scaling the CIRES-20CR PC2 time series so that its standard deviation and mean are identical to that from the AVHRR data over their common time period of 1982–2009 (blue line in Fig. 3a). The CIRES-20CR PC2 time series is highly correlated with the ERA-I PC2 series (orange line in Fig. 3a) at an $r$ value of 0.63 ($P < 0.01$), the AVHRR dust optical depth time series (yellow line in Fig. 3a) at $r = 0.54$ ($P < 0.02$), and a 54-year dust proxy dust time series based on AVHRR data and other data from a Cape Verde coral[2] (purple line) at $r = 0.55$ ($P < 0.01$). These $r$ and $P$ values were calculated using unsmoothed annual time series.

The positive and statistically significant correlations between the CIRES-20CR PC2 time series and the other dust time series in Fig. 3a are evidence that this record of over 150 years can be used to study historical North African dust emission and transport. Over the entire CIRES-20CR PC2 time series, there are two relatively persistent periods of increased dustiness, 1910–1950 and 1970–1990, and three periods of low dust concentrations, the mid-1860s to early 1870s, 1950 to the late 1960s, and the last 15 years of the record. There is not a secular trend in dust emission and transport over the entire record, but there are coherent multi-decadal trend periods, including upward trends from 1870–1910 and the late 1950s to the mid-1980s, and downward trends from the mid-1940s to 1960 and the mid-1980s to the end of the record. In addition, the variance of the first half of the CIRES-20CR PC2 time series ($4.9 \times 10^{-3}$) is 58% of the variance of the second half of the record ($8.4 \times 10^{-3}$), raising the possibility that anthropogenic forcing has caused an enhancement in year-to-year changes in the amount of dust emitted from North Africa.

Our results suggest that any phenomenon that excites surface winds over North Africa in a manner that projects onto the spatial structure in Fig. 1a will influence the net production and transport of Saharan dust. Such a theory reconciles the disparity of published work on the controls on North African dust production and suggests that, to first order, to understand dust variability one needs only to elucidate the factors that excite the pattern of surface wind speeds in Fig. 1a. Our results explain why although time series of dust are correlated with a diversity of climate phenomena, these correlations are not stationary in time. For example, an analysis of correlation coefficients between different climate indices and CIRES-20CR PC2, for a moving 31-year window, suggests that the high dust emissions of the 1910s and 1940s were related to the phase of the North Atlantic Oscillation, but that the high dust emissions of the 1980s were associated with the Sahelian drought (Extended Data Fig. 4).

We next examine simulated changes in African dust during the twenty-first century using output from the Fifth Climate Model Intercomparison Project (CMIP5, Extended Data Table 1). As CMIP5 models are unable to reproduce observed twentieth-century changes in African dust[22] we estimate future dustiness by projecting the EOF pattern in Fig. 1 onto the monthly mean 10-m wind fields for these models. We first consider winds from the so-called 'business-as-usual' scenario (RCP 8.5), in which emissions of greenhouse gases continue to increase throughout the twenty-first century[23]. We convert ensemble mean CMIP5 PC2 time series to unit dust optical depth by scaling each series so that its standard deviation is equal to the standard deviation of the entire CIRES-20CR PC2 time series (Fig. 3b).

We also offset each model's mean optical depth so that the mean of the first five model years (2005–2010) is equal to 0.4, the long-term mean of the CIRES-20CR PC2 time series (blue dashed line, Fig. 3b).

The twenty-first-century CMIP5 PC2 time series for each model is a typical 'spaghetti plot' (grey lines in Fig. 3b), reflecting the models' internal variability; the range of individual model dust optical depth values is 0.2 to 0.6. The CMIP5 multimodel mean time series (thick black line) shows a statistically significant downward trend of $-0.04 \pm 0.01$ (in units of dust optical depth) per 100 years (red line in Fig. 3b). This 100-year change is approximately 40% of the magnitude of the models' standard deviation over the same period and represents a 10% reduction in the CIRES-20CR PC2 long-term mean dust optical depth. Furthermore, 15 of the 34 models considered here have statistically significant downward trends in dust optical depth yet only six models have statistically significant upward trends in dust optical depth (Extended Data Fig. 5a). We repeated this analysis for the RCP 4.5 simulations, in which emissions of greenhouse gases peak during 2040–2050, finding a statistically significant downward trend in dust optical depth of $-0.02 \pm 0.01$, half of that for the RCP 8.5 simulation (Extended Data Fig. 5b). The multimodel mean time series from the RCP scenarios represent an estimate of the response of dust to increasing levels of greenhouse gasses and do not reflect other sources of variability.

From our analysis of the CMIP5 data we conclude that the future reduction in emission and transport of dust from Africa is a robust response to increasing greenhouse gas emissions. The CMIP5 trends were not sensitive to the types of aerosol indirect effects included in the model, suggesting that the trend is not related to twenty-first-century changes in the concentration of atmospheric aerosols included in the simulations. We estimate the 'time of emergence' of the greenhouse gas forced trend to be 200 years ($P = 0.05$) based on bootstrap resampling tests using noise characteristics from the CIRES-20CR PC2 time series (see Methods), and thus the trend in the CMIP5 multimodel mean is consistent with the lack of a trend over the entire CIRES-20CR PC2 time series (Fig. 3). However, there is a statistically significant downward trend in CIRES-20CR PC2 over the twentieth century ($-0.08 \pm 0.06$ per 100 years). It is plausible that the twentieth- and twenty-first-century PC2 downward trends are associated with a slow-down of the tropical circulation[24].

This decline in dust over North Africa may result in a slight improvement in air quality in the region, although the effect of regional population growth and urbanization will undoubtedly overshadow the benefits of a reduction in airborne dust[25]. While the radiative forcing of dust may be near zero over North Africa, as shortwave cooling is approximately balanced by the longwave warming[6], dust transported over the tropical North Atlantic cools the surface via direct[2,15] and indirect[26] radiative effects. Therefore, a reduction in dust would act as a positive feedback to warming by greenhouse gases in the tropical North Atlantic. Furthermore, since this feedback is not pan-tropical, this additional dust-forced warming could increase hurricane activity by increasing tropical North Atlantic sea surface temperature[27], relative sea surface temperature (which is the difference between sea surface temperature in the tropical Atlantic and the sea surface temperature averaged over all of the tropics[28]), and the northward meridional sea surface temperature gradient[4,29]. The radiative and temperature effects from such a reduction in dust are not captured in most CMIP5 simulations; many models do not have interactive dust, and of those models that do, the majority show an increase in simulated African dust concentrations during the twenty-first century (Extended Data Fig. 6). Thus, it is plausible that current temperature projections for the tropical Atlantic through the Caribbean are too conservative.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 5 October 2015; accepted 11 January 2016.**

1. Chaboureau, J. P., Tulet, P. & Mari, C. Diurnal cycle of dust and cirrus over West Africa as seen from Meteosat Second Generation satellite and a regional forecast model. *Geophys. Res. Lett.* **34,** L02822 (2007).
2. Evan, A. T. & Mukhopadhyay, S. African dust over the northern tropical Atlantic: 1955–2008. *J. Appl. Meteorol. Climatol.* **49,** 2213–2229 (2010).
3. Bristow, C. S., Hudson-Edwards, K. A. & Chappell, A. Fertilizing the Amazon and equatorial Atlantic with West African dust. *Geophys. Res. Lett.* **37,** L14807 (2010).
4. Evan, A. T., Foltz, G. R., Zhang, D. & Vimont, D. J. Influence of African dust on ocean-atmosphere variability in the tropical Atlantic. *Nature Geosci.* **4,** 762–765 (2011).
5. Ridley, D. A., Heald, C. L. & Prospero, J. M. What controls the recent changes in African mineral dust aerosol across the Atlantic? *Atmos. Chem. Phys.* **14,** 5735–5747 (2014).
6. Yoshioka, M. *et al.* Impact of desert dust radiative forcing on Sahel precipitation: relative importance of dust compared to sea surface temperature variations, vegetation changes, and greenhouse gas warming. *J. Clim.* **20,** 1445–1467 (2007).
7. DeFlorio, M. J. *et al.* Interannual modulation of subtropical Atlantic boreal summer dust variability by ENSO. *Clim. Dyn.* **46,** 585–599 (2015).
8. Prospero, J. M. & Lamb, P. J. African droughts and dust transport to the Caribbean: climate change implications. *Science* **302,** 1024–1027 (2003).
9. Moulin, C. *et al.* Control of atmospheric export of dust from North Africa by the North Atlantic Oscillation. *Nature* **287,** 691–694 (1997).
10. Doherty O. M., Riemer N. & Hameed S. Control of Saharan mineral dust transport to Barbados in winter by the Intertropical Convergence Zone over West Africa. *J. Geophys. Res.* **117,** D19117 (2012).
11. Doherty, O. M., Riemer, N. & Hameed, S. Role of the convergence zone over West Africa in controlling Saharan mineral dust load and transport in the boreal summer. *Tellus B* **66,** 23191 (2014).
12. Wang, W. J., Evan, A. T., Flamant, C. & Lavaysse, C. On the decadal scale correlation between African dust and Sahel rainfall: the role of Saharan heat low-forced winds. *Science Adv.* **1,** e1500646 (2015).
13. Compo, G. P. *et al.* The twentieth century reanalysis project. *Q. J. R. Meteorol. Soc.* **137,** 1–28 (2011).
14. Griffin, D. W. & Kellogg, C. A. Dust storms and their impact on ocean and human health: dust in Earth's atmosphere. *EcoHealth* **1,** 284–295 (2004).
15. Evan, A. T., Vimont, D. J., Heidinger, A. K., Kossin, J. P. & Bennartz, R. The role of aerosols in the evolution of tropical North Atlantic ocean temperature anomalies. *Science* **324,** 778–781 (2009).
16. Dunion, J. P. & Velden, C. S. The impact of the Saharan air layer on Atlantic tropical cyclone activity. *Bull. Am. Meteorol. Soc.* **85,** 353–365 (2004).
17. Dee, D. P. *et al.* The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137,** 553–597 (2011).
18. Washington, R. *et al.* Links between topography, wind, deflation, lakes and dust: the case of the Bodélé depression, Chad. *Geophys. Res. Lett.* **33,** L09401 (2006).
19. Prospero, J. M., Ginoux, P., Torres, O., Nicholson, S. E. & Gill, T. E. Environmental characterization of global sources of atmospheric soil dust identified with the Nimbus 7 Total Ozone Mapping Spectrometer (TOMS) absorbing aerosol product. *Rev. Geophys.* **40,** 1002 (2002).
20. Evan, A. T. *et al.* Derivation of an observation-based map of North African dust emission. *Aeolian Res.* **16,** 153–162 (2015).
21. Compo, G. P. *et al.* The twentieth century reanalysis project. *Q. J. R. Meteorol. Soc.* **137,** 1–28 (2011).
22. Evan, A. T., Flamant, C., Fiedler, S. & Doherty, O. An analysis of aeolian dust in climate models. *Geophys. Res. Lett.* **41,** 5996–6001 (2014).
23. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93,** 485–498 (2012).
24. Held, I. M. & Soden, B. J. Robust responses of the hydrological cycle to global warming. *J. Clim.* **19,** 5686–5699 (2006).
25. Liousse, C. *et al.* Explosive growth in African combustion emissions from 2005 to 2030. *Environ. Res. Lett.* **9,** 035003 (2014).
26. Doherty, O. M. & Evan, A. T. Identification of a new dust-stratocumulus indirect effect over the tropical North Atlantic. *Geophys. Res. Lett.* **41,** 6935–6942 (2014).
27. Emanuel, K. A. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature* **436,** 686–688 (2005).
28. Vecchi, G. A. & Soden, B. J. Effect of remote sea surface temperature change on tropical cyclone potential intensity. *Nature* **450,** 1066–1070 (2007).
29. Kossin, J. P. & Vimont, D. J. A more general framework for understanding Atlantic hurricane variability and trends. *Bull. Am. Meteorol. Soc.* **88,** 1767–1781 (2007).

**Acknowledgements** This work was supported by the Agence Nationale de la Recherche (ANR) grant ANR-10-LABX-18-01 of the national Programme Investissements d'Avenir provided by the Laboratoire d'excellence Institut Pierre Simon Laplace (L-IPSL). We appreciate comments from our colleagues.

**Author Contributions** A.T.E. carried out the main analysis and wrote the manuscript. F.G. analysed wind speed data from weather stations. All authors designed the study, discussed the results and commented on the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.T.E. (aevan@ucsd.edu).

## METHODS

**EOF and PC calculations.** To calculate the EOFs of 10-m winds over North Africa we removed the mean and seasonal cycle from the monthly mean zonal and meridional wind fields from ERA-I for the period 1979–2014. ERA-I contains global atmospheric parameters from January 1979 to the present, at T255 spectral resolution (approximately 80 km). We spatially smoothed the monthly mean data via a uniformly weighted 3° × 3° filter and divided the data into subsets from −20° to 30° E and 2° to 35° N, with over-water values masked out. Eigenvectors were calculated from these smoothed and subsetted fields. The PC time series were calculated by projecting the subsequent eigenvalues back onto the data, and the EOF spatial pattern (Fig. 1a) is the regression of the meridional and zonal wind fields onto the PC time series. The PC time series for the CIRES-20CR (Fig. 3a) was calculated by smoothing and subsetting the CIRES-20CR 10-m zonal and meridional wind fields in a manner identical to that done for the ERA-I data. We then projected the second eigenvector from the ERA-I EOF analysis (Fig. 1) onto the CIRES-20CR wind fields to derive an equivalent PC time series for the CIRES-20CR data set.

We repeated the eigenanalysis of 10-m winds using other reanalysis products, including the NOAA-CIRES Twentieth-Century Reanalysis[13], the ERA Twentieth-Century Reanalysis[31], the NASA Modern-Era Retrospective analysis for Research and Applications[32], the NCEP-DOE AMIP-II Reanalysis[33] and the NCEP NCAR Reanalysis[34] (Extended Data Fig. 7). While the results from the eigenanalysis of these other reanalysis products show elements of the second EOF from ERA-I, these were mixed among the first two or three EOFs, and rotation of the EOFs did not clearly separate out the dust signal as is seen in the ERA-I data (Fig. 1b).

Recent work has shown that, when compared to observations of surface winds from meteorological stations across the Sahel, 10-m winds from ERA-I are more accurate than those from other reanalyses[30]. We expanded on the analysis in ref. 30 to include stations in the Sahara, also finding that here 10-m winds from ERA-I were more accurate than those from other reanalysis products (Extended Data Fig. 8).

We also examined the PC2 time series from the CMIP5 historical forcing experiments. These CMIP5 PC2 time series for individual models are constructed in a manner identical to that for the RCP 8.5 simulations in Fig. 3. The resultant multimodel mean time series had a standard deviation approximately 20% of that from the CIRES-20CR, and thus to facilitate comparison between the two we scaled the multimodel mean time series to have a standard deviation equal to that of the CIRES-20CR time series (Extended Data Fig. 9). The CIRES-20CR and the CMIP5 multimodel mean show some agreement in their trends, including a rise in dust emission and transport from 1880 to 1930 and a reduction in dust over the end of the record.

Variability in the multimodel mean historical forcing simulations should reflect external forcing only, that is, associated with variations in solar insolation, greenhouse gases or aerosol concentrations but not internal variability of the climate system. Thus, disagreement between the two time series may be due to a combination of internal variability of the physical climate system and poor representation of key processes controlling surface winds over North Africa on these timescales, where the latter may also be a major reason why the variance in the multimodel mean time series was one-fifth of that from CIRES-PC2.

**Linearization of the relationship between wind speed and dust emission.** The results in Fig. 1b suggest that monthly mean dust emission and transport is, to first order, linearly proportional to monthly mean surface wind speed. Such an assumption is common, although often implicit, when examining the interannual variability of dust (for example, refs 9–12), which may be unexpected given that dust emission is proportional to the cube of wind speed (for example, ref. 35). To test this assumption we analysed hourly data from three synoptic stations in the Sahara that are close to the major dust source regions (Djanet and Tamanrasset in Algeria and Agadez in Niger). We calculated the so-called dust uplift potential (DUP)[36] using the characteristic emission threshold of 6 m s$^{-1}$ (refs 37–39). From the hourly data we calculated the monthly mean wind speed and the monthly mean DUP. At each station DUP is highly correlated with wind speed, with $r$ values ranging from 0.86 (Djanet) to 0.93 (Agadez) and $P$ values $< 0.01$ (Extended Data Fig. 2). These results demonstrate that on monthly and longer timescales the monthly mean dust emission is, to first-order, proportional to monthly mean wind speed.

As another test of the linearity assumption, and to remove potential biases related to seasonality of series autocorrelation, for each station we randomly drew 24 × 30 wind speed samples from the data set (representing a month of hourly wind speed observations), calculated the DUP, and then averaged these DUP and wind speed values. We repeated this procedure 100 times and calculated the correlation between the randomly sampled wind speed and DUP values, finding qualitatively similar $r$ and $P$ values for correlations between this monthly mean DUP and the measured wind speed at each station (not shown). We repeated this procedure, increasing the sampling to 1000 and 10,000, obtaining similar results with both (not shown). As a final test of the linearity assumption, we fitted the wind speed distribution for each station to a lognormal distribution, randomly sampled then averaged the distributions in a manner identical to that described above, and then calculated the pseudo monthly mean values of DUP and wind speed. The resultant correlation coefficients between DUP and wind speed were also qualitatively similar (not shown).

Although local changes in soil moisture and vegetation do influence dust emission[37,40], our results suggest that variability in North African dust emission and transport is little affected by such properties of the surface. This is because the main dust source regions (Fig. 2) are within the hyper-arid Sahara (defined as a mean annual rainfall of less than 100 mm), where soil moisture and vegetation are extremely limited (for example, ref. 41). The negligible influence of soil moisture and vegetation to the interannual variability of North African dust is consistent with recent modelling work[5] as well as previous studies on the factors governing the interannual variability of dust (for example, refs 9–12).

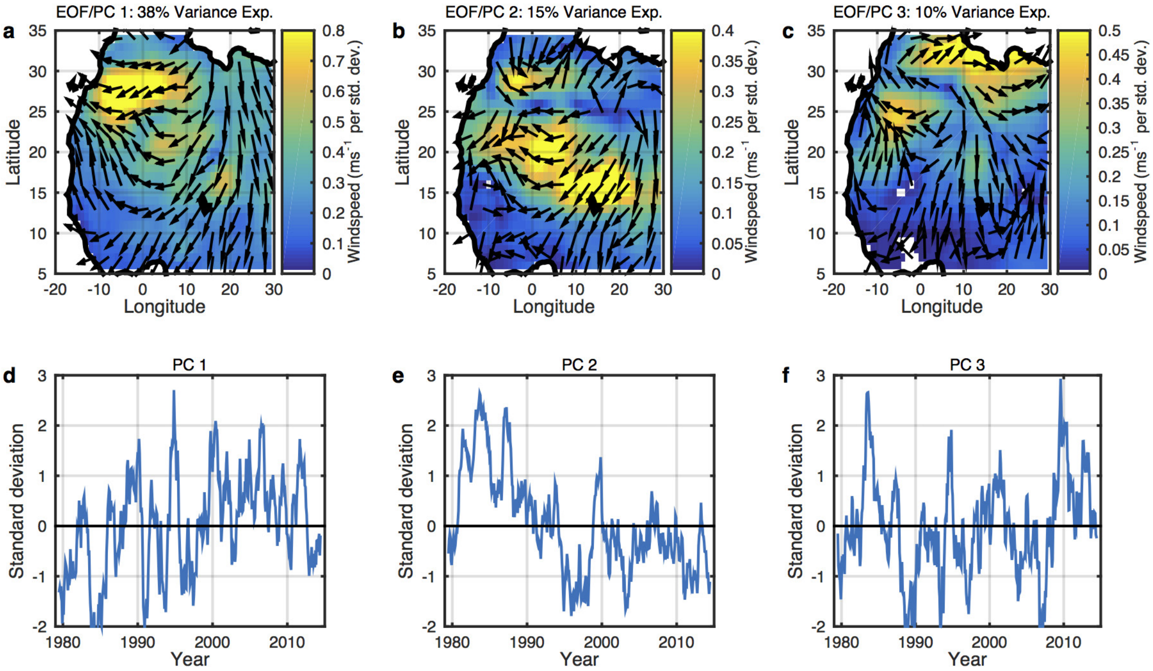
**Time of emergence.** The time of emergence of the PC2 trends is an estimate of the length of the time series required to detect a trend at 95% confidence level, which we determined via a bootstrap method. To do so we created 1000 random (normal) time series of length 500 years with standard deviation equal to that of the CIRES-20CR PC2 time series (Fig. 3a) and a linear trend equal to that of the RCP 8.5 multimodel mean (Extended Data Fig. 5a). We calculated the linear least-squares trend and the 95% confidence interval on the trend for the time series using the first 2 years, then the first 3 years, then the first 4 years, through to 500 years. The time of emergence is then calculated as the length of the time series for which the trend is statistically different from zero at the 95% confidence interval, which in this case is 195 years.

**Data.** CMIP5 data are available through the Earth System Grid (http://pcmdi9.llnl.gov/). Satellite and proxy dust data are available through the PANGAEA Data Publisher for Earth and Environmental Science (http://doi.pangaea.de/10.1594/PANGAEA.855141) and a map of North African dust emission is available from the same resource at http://doi.pangaea.de/10.1594/PANGAEA.855243. ERA-I and ERA Twentieth-Century Reanalysis data are available from the European Centre for Medium-Range Weather Forecasts (ERA-20CR) (http://www.ecmwf.int/). CIRES-20CR, DOE AMIP-II (NCEP2) and NCEP NCAR (NNRP) Reanalysis data are from the Earth System Research Laboratory (http://www.esrl.noaa.gov/). NASA Modern-Era Retrospective analysis for Research and Applications (MERRA) data are from the Goddard Earth Sciences Data and Information Services Center (http://disc.sci.gsfc.nasa.gov). Wind speed data from Saharan synoptic weather stations are from the Wyoming Weather Web (http://weather.uwyo.edu) and the AMMA database (http://database.amma-international.org). The Niño 3.4 time series and Palmer Drought[42,43] data are available from the NOAA Climate Prediction Center (http://www.cpc.ncep.noaa.gov). Jones North Atlantic Oscillation[44] data are from the UK Climate Research Unit (http://www.cru.uea.ac.uk/). The Saharan Heat Low thickness time series is calculated using ERA-I data via the methodology in ref. 45 and data on the latitude of the intertropical convergence zone are calculated using NNRP data via the methodology in ref. 10.

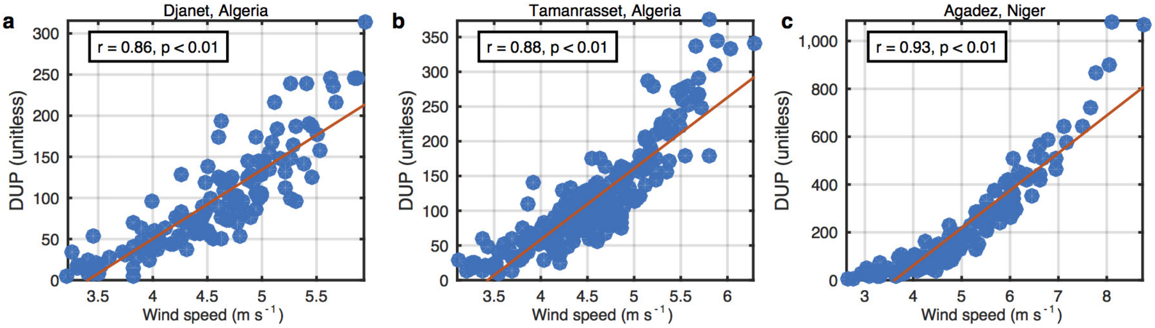
**Code availability.** The codes used to conduct the analysis presented in this paper and in the production of the figures are available at https://github.com/amatoevan/2015DUST/.

30. Largeron, Y. *et al.* Can we use surface wind fields from meteorological reanalyses for Sahelian dust emission simulations? *Geophys. Res. Lett.* **42,** 2490–2499 (2015).
31. Dee, D., Balmaseda, D. M., Balsamo, G., Engelen, R. & Simmons, A. Toward a consistent reanalysis of the climate system. *Bull. Am. Meteorol. Soc.* **95,** 1235–1248 (2014).
32. Rienecker, M. M. *et al.* MERRA: NASA's modern-era retrospective analysis for research and applications. *J. Clim.* **24,** 3624–3648 (2011).
33. Kanamitsu, M. *et al.* NCEP-DOE AMIP-II reanalysis (R-2). *Bull. Am. Meteorol. Soc.* **83,** 1631–1643 (2002).
34. Kalnay, E. *et al.* The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77,** 437–471 (1996).
35. Marticorena, B. & Bergametti, G. Modeling the atmospheric dust cycle: 1. Design of a soil-derived dust emission scheme. *J. Geophys. Res.* **100,** 16415–16430 (1995).
36. Marsham, J. H., Knippertz, P., Dixon, N. S., Parker, D. J. & Lister, G. M. S. The importance of the representation of deep convection for modeled dust-generating winds over West Africa during summer. *Geophys. Res. Lett.* **38,** L16803 (2011).
37. Cowie, S. M., Knippertz, P. & Marsham, J. H. Are vegetation-related roughness changes the cause of the recent decrease in dust emission from the Sahel? *Geophys. Res. Lett.* **40,** 1868–1872 (2013).

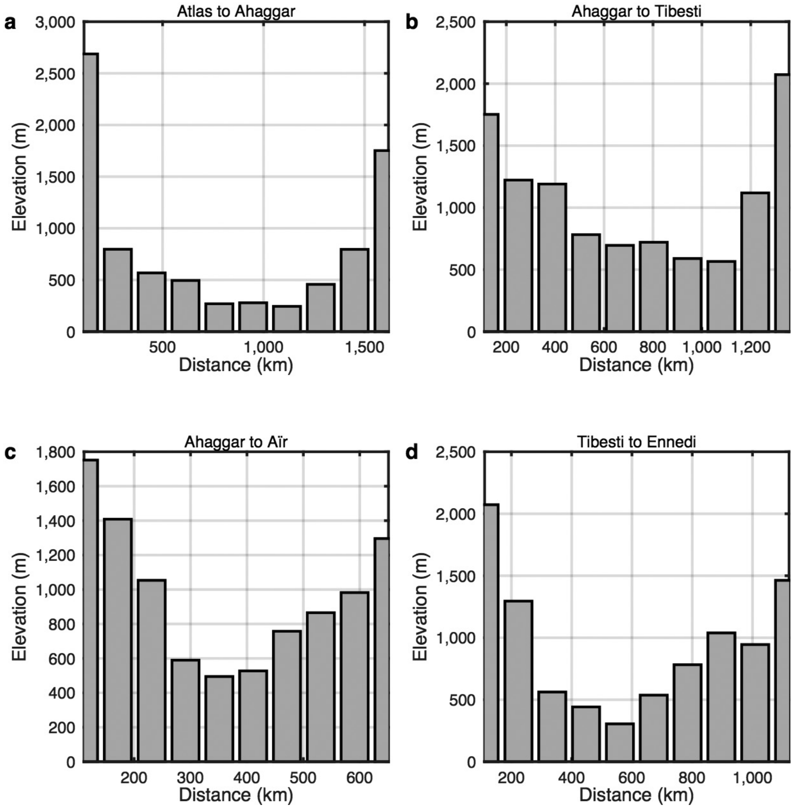
38. Westphal, D. L., Toon, O. B. & Carlson, T. N. A two-dimensional numerical investigation of the dynamics and microphysics of Saharan dust storms. *J. Geophys. Res.* **92,** 3027–3049 (1987).
39. Tegen, I. & Fung, I. Modeling of mineral dust in the atmosphere: sources, transport, and optical thickness. *J. Geophys. Res.* **99,** 22897–22914 (1994).
40. Fécan, F., Marticorena, B. & Bergametti, G. Parameterization of the increase of the aeolian erosion threshold wind friction velocity due to soil moisture for arid and semi-arid areas. *Ann. Geophys.* **17,** 149–157 (1999).
41. Washington, R., Todd, M. C., Engelstaedter, S., Mbainayel, S. & Mitchell, F. Dust and the low-level circulation over the Bodélé Depression, Chad: observations from BoDEx 2005. *J. Geophys. Res.* **111** (D3), D03201 (2006).
42. Dai, A., Trenberth, K. E. & Qian, T. A global data set of Palmer Drought Severity Index for 1870-2002: relationship with soil moisture and effects of surface warming. *J. Hydrometeorol.* **5,** 1117–1130 (2004).
43. Mahowald, N. M. *et al.* Observed 20th century desert dust variability: impact on climate and biogeochemistry. *Atmos. Chem. Phys.* **10,** 10875–10893 (2010).
44. Jones, P. D., Jonsson, T. & Wheeler, D. Extension to the North Atlantic Oscillation using early instrumental pressure observations from Gibraltar and South-West Iceland. *Int. J. Climatol.* **17,** 1433–1450 (1997).
45. Lavaysse, C. *et al.* Seasonal evolution of the West African heat low: a climatological perspective. *Clim. Dyn.* **33,** 313–330 (2009).



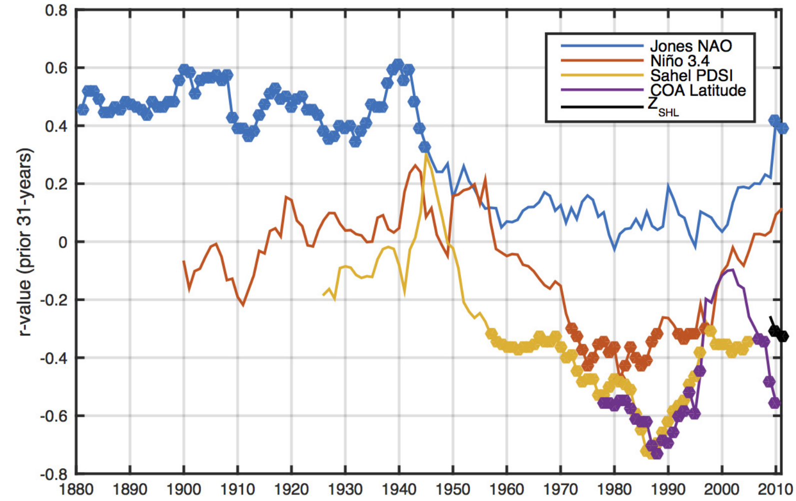
**Extended Data Figure 1 | First three EOFs and PCs of 10-m winds over North Africa from ERA-I.** Shown is the spatial structure of the first three EOFs from the eigenanalysis of monthly mean 10-m winds from ERA-I (**a**–**c**), and the corresponding PC time series in units of standard deviation (**d**–**f**). Descriptions of arrows, shading and time series are identical to that for Fig. 1. See Methods for details of the eigenanalysis.



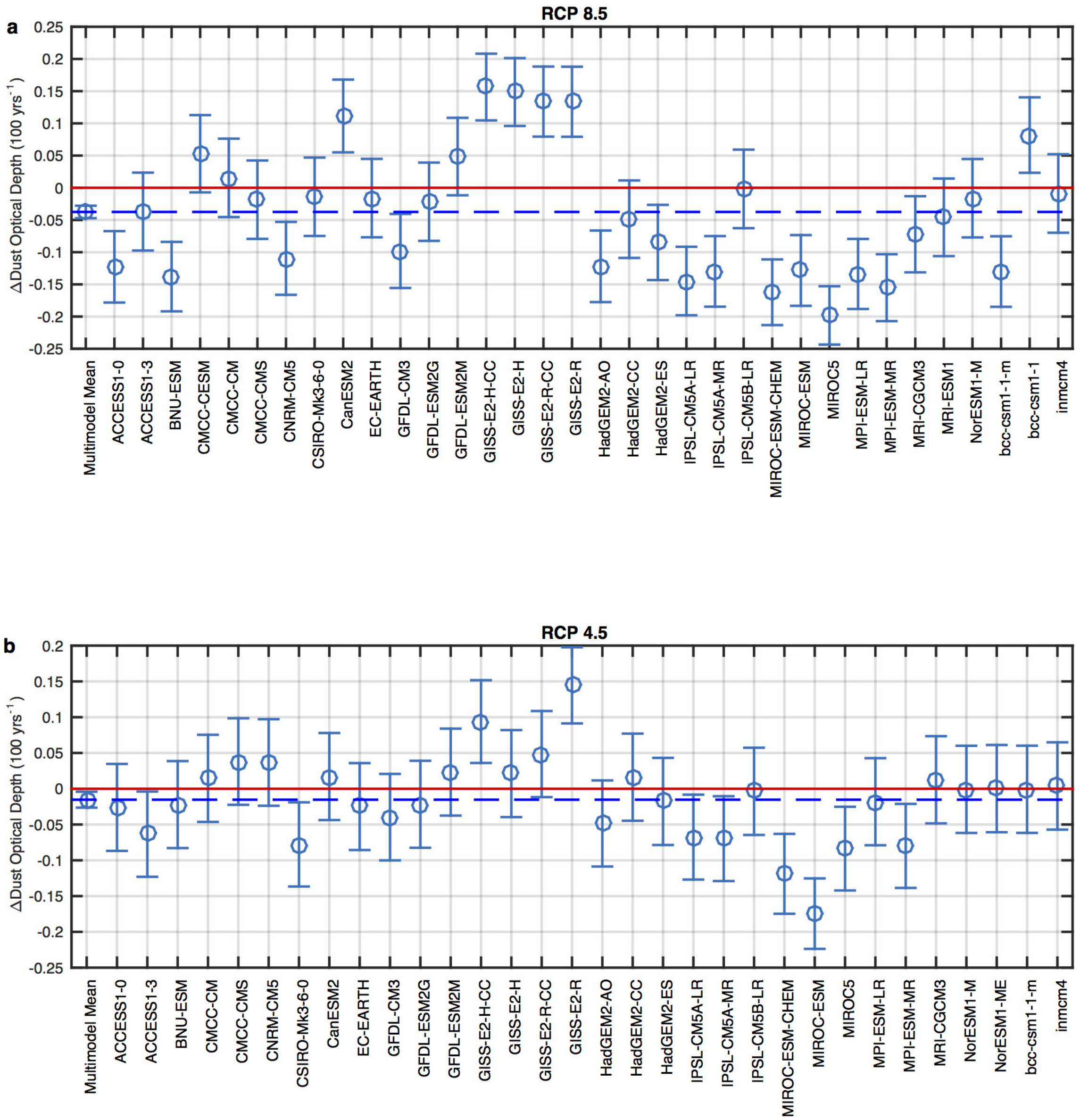
**Extended Data Figure 2 | Comparison of monthly mean wind speeds and dust uplift potential.** Shown are scatter plots (blue filled circles) of mean monthly DUP (ordinate axis) and wind speeds (abscissa axis) for the North African synoptic stations at Djanet, Algeria (**a**), Tamanrasset, Algeria (**b**) and Agadez, Niger (**c**). Also shown for reference are the least-squares best-fit lines (red). The correlation coefficient *r* and statistical significance *P* are indicated in each plot.



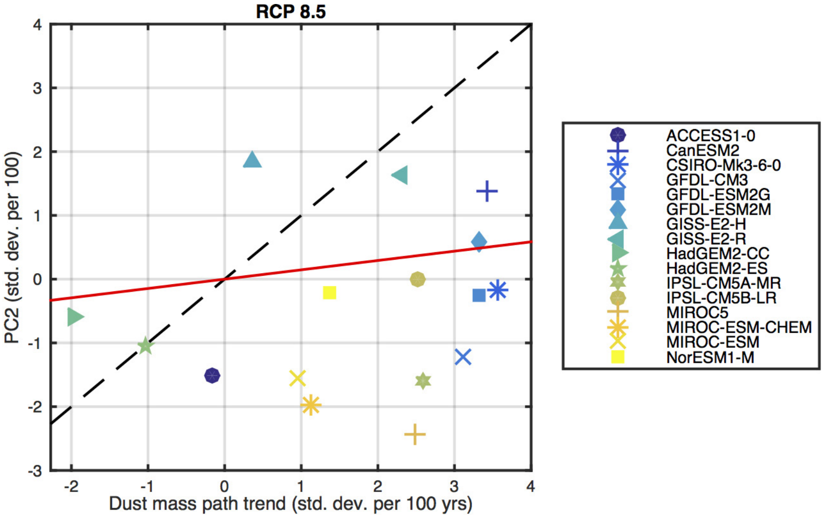
**Extended Data Figure 3 | Topography affecting surface winds across the Sahara.** Shown are surface elevations along each transect in Fig. 2: the Atlas to the Ahaggar mountains (**a**), the Ahaggar to the Tibesti mountains (**b**), the Ahaggar mountains to the Aïr massif (**c**) and the Tibesti mountains to the Ennedi plateau (**d**). These topographic features accelerate the surface flow and give rise to the spatial structure of the second EOF and PC pair (Fig. 1a). The titles indicate the major topographic features bounding each transect. Note the different horizontal lengths (abscissa) and heights (ordinate) for each plot.



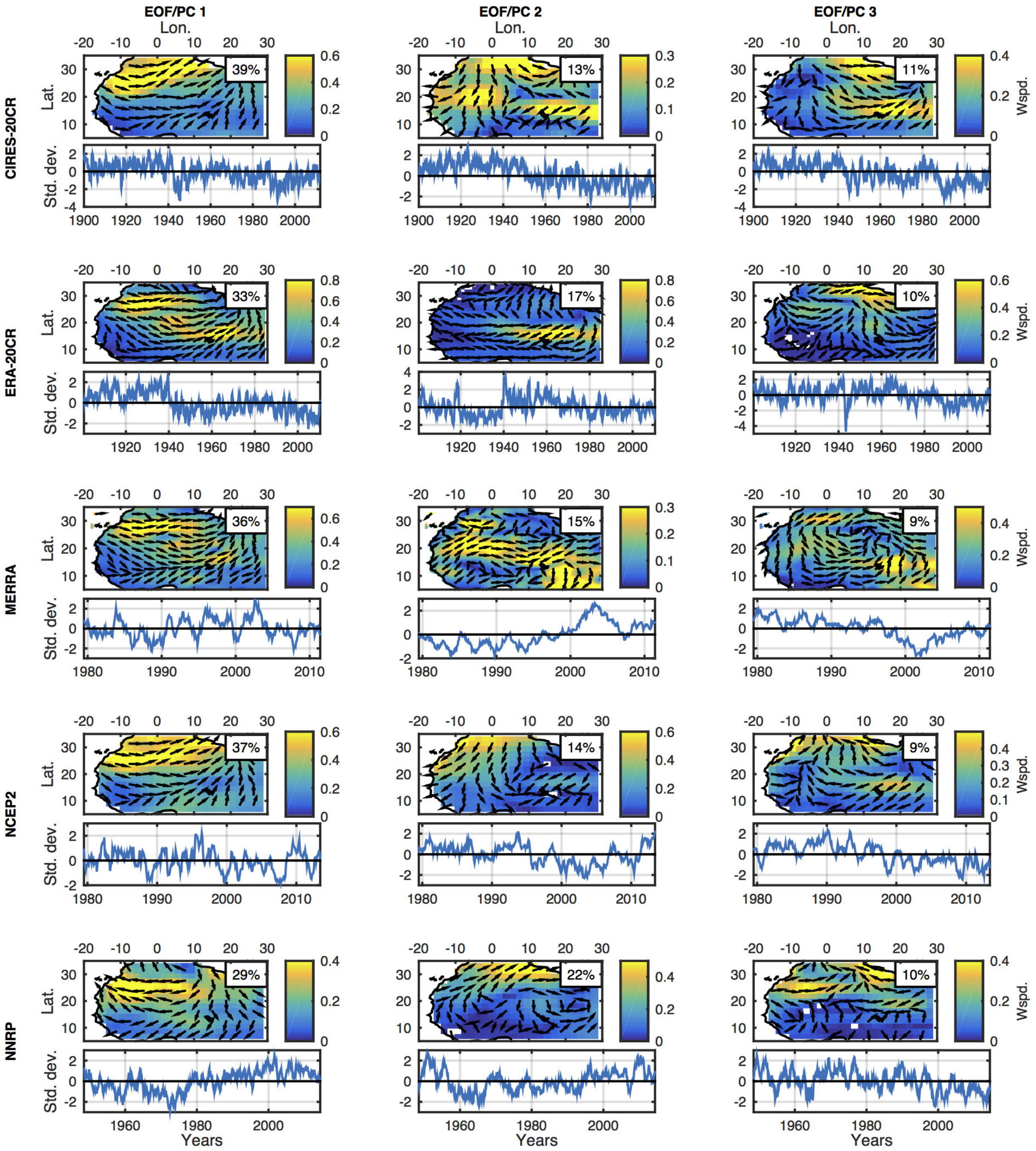
**Extended Data Figure 4 | Non-stationary correlations between CIRES-20CR PC2 and climate indices.** Plotted is the correlation coefficient between the annual mean CIRES-20CR PC2 and the Jones North Atlantic Oscillation, Niño 3.4, the Sahel-averaged Palmer Drought Severity index (PDSI), the latitude of the intertropical convergence zone, and the 925–700 hPa thickness of the Saharan Heat Low ($Z_{SHL}$). All correlation coefficients are for the preceding 31-year period (for example, the value of 0.6 for the Jones North Atlantic Oscillation in 1940 indicates that the correlation coefficient between the Jones North Atlantic Oscillation and the CIRES-20CR PC2 is 0.6 for the period 1910–1940). We indicate statistically significant correlations ($P < 0.05$) with a filled circle, although here the $P$ value is not calculated using effective degrees of freedom (as is the case elsewhere in this Letter).



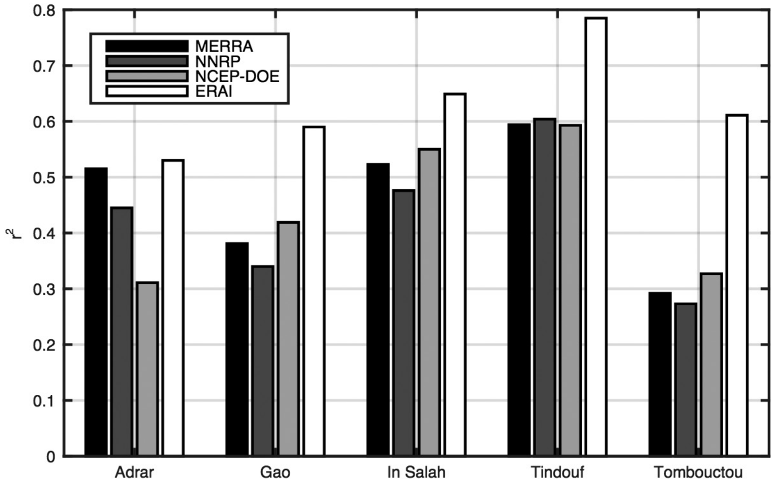
**Extended Data Figure 5 | CMIP5 RCP 4.5 and RCP 8.5 twenty-first-century trends in PC2. a**, RCP 8.5. **b**, RCP 4.5. Shown are the PC2 linear trends (circles), 95% confidence intervals (error bars) and the multimodel mean trend (blue dashed line) for RCP 8.5 (top) and 4.5 (bottom) simulations. All trends are in units of dust optical depth per 100 years.



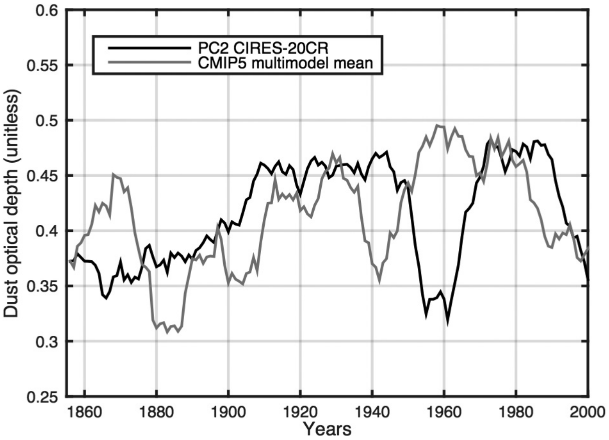
**Extended Data Figure 6 | CMIP5 RCP 8.5 twenty-first-century trends in estimated and modelled dust.** Plotted is the CMIP5 models' twentieth-century trends in dust mass path for the RCP 8.5 experiments (abscissa) versus the twentieth-century trends calculated from the PC2 time series (ordinate). Included here are only CMIP5 models for which dust mass path and 10-m wind data are available. The red line is the least-squares best-fit line (slope is not statistically different from zero) and the dashed line is the one-to-one line. All trends are in units of standard deviation per 100 years.



**Extended Data Figure 7 | EOF/PC pairs for various reanalyses.** Shown are the first three EOFs (top rows) and corresponding PC time series (bottom rows) calculated from the CIRES-20CR, ERA-20CR, MERRA, NCEP2 and NNRP data sets. Percentages of the variance of the data set explained by each EOF/PC pair are shown.



**Extended Data Figure 8 | Wind speed comparisons between observations and reanalyses.** Shown are the $r^2$ values from the correlation between monthly mean surface winds from station data and monthly mean 10-m winds from reanalyses for five stations in the Sahara over the period 2000–2013. In all cases the $r^2$ values for ERA-I are greater than 0.5 and higher than those for the other reanalyses.



**Extended Data Figure 9 | Modelled and reanalysis time series of historical dust.** Plotted is the PC2 time series from the CIRES-20CR, identical to that shown in Fig. 3a, and the CMIP5 multimodel mean PC2 time series from the historical forcing simulations. Both annul mean time series have been smoothed with an 11-point running mean filter to highlight decadal scale variability.

**Extended Data Table 1 | CMIP5 models used in this study**

| Institution | Model | Historical | RCP4.5 | RCP8.5 |
|---|---|---|---|---|
| Commonwealth Scientific and Industrial Research Organisation (CSIRO) and Bureau of Meteorology (BOM; Australia) | ACCESS1.0 | 3 | 1 | 1 |
| | ACCESS1.3 | 3 | 1 | 1 |
| Beijing Climate Center (BCC; China) | BCC-CSM1.1 | 0 | 0 | 1 |
| | BCC-CSM1.1(m) | 0 | 1 | 1 |
| Global Change and Earth System Science (GCESS), Beijing Normal University (BNU; China) | BNU-ESM | 1 | 1 | 1 |
| Canadian Centre for Climate Modelling and Analysis (CCCma; Canada) | CanESM2 | 5 | 5 | 5 |
| Centro Euro-Mediterraneo per I Cambiamenti Climatici (CMCC; Italy) | CMCC-CESM | 1 | 0 | 1 |
| | CMCC-CM | 1 | 1 | 1 |
| | CMCC-CMS | 1 | 1 | 1 |
| Centre National de Recherches Météorologiques (CNRM)–Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS; France) | CNRM-CM5 | 9 | 1 | 4 |
| CSIRO–Queensland Climate Change Centre of Excellence (QCCE; Australia) | CSIRO-Mk3.6.0 | 0 | 9 | 9 |
| European consortium (EC) | EC-EARTH | 1 | 3 | 2 |
| National Oceanographic and Atmospheric Administration (NOAA) Geophysical Fluid Dynamics Laboratory (GFDL; United Stated) | GFDL-CM3 | 0 | 1 | 1 |
| | GFDL-ESM2G | 0 | 0 | 1 |
| | GFDL-ESM2M | 0 | 1 | 1 |
| National Aeronautics and Space Administration (NASA) Goddard Institute for Space Studies (GISS; United Stated) | GISS-E2-H-CC | 0 | 1 | 1 |
| | GISS-E2-H | 6 | 5 | 2 |
| | GISS-E2-R-CC | 0 | 1 | 1 |
| | GISS-E2-R | 6 | 6 | 2 |
| National Institute of Meteorological Research (NIMR)–Korea Meteorological Administration (KMA; South Korea) | HadGEM2-AO | 0 | 1 | 1 |
| Met Office Hadley Centre (MOHC; United Kingdom) | HadGEM2-CC | 0 | 1 | 1 |
| | HadGEM2-ES | 0 | 4 | 4 |
| Institute for Numerical Mathematics (INM; Russia) | INM-CM4 | 1 | 1 | 1 |
| L'Institut Pierre-Simon Laplace (IPSL; France) | IPSL-CM5A-LR | 6 | 3 | 3 |
| | IPSL-CM5A-MR | 3 | 1 | 1 |
| | IPSL-CM5B-LR | 1 | 1 | 1 |
| Model for Interdisciplinary Research on Climate (MIROC; Japan) | MIROC5 | 0 | 3 | 3 |
| | MIROC-ESM | 3 | 1 | 1 |
| | MIROC-ESM-CHEM | 1 | 1 | 1 |
| Max Planck Institute for Meteorology (MPI-M; Germany) | MPI-ESM-LR | 3 | 3 | 3 |
| | MPI-ESM-MR | 3 | 3 | 1 |
| Meteorological Research Institute (MRI; Japan) | MRI-CGCM3 | 5 | 1 | 1 |
| | MRI-ESM1 | 0 | 0 | 1 |
| Norwegian Climate Centre (NCC; Norway) | NorESM1-M | 3 | 1 | 1 |
| | NorESM1-ME | 1 | 1 | 0 |

Shown are the modelling centres, model names, and the number of ensemble members for both RCP experiments and the historical forcing experiment examined in this study.

doi:10.1038/nature17160

# Intrinsic honesty and the prevalence of rule violations across societies

Simon Gächter[1,2,3] & Jonathan F. Schulz[1,4]

**Deception is common in nature and humans are no exception[1]. Modern societies have created institutions to control cheating, but many situations remain where only intrinsic honesty keeps people from cheating and violating rules. Psychological[2], sociological[3] and economic theories[4] suggest causal pathways to explain how the prevalence of rule violations in people's social environment, such as corruption, tax evasion or political fraud, can compromise individual intrinsic honesty. Here we present cross-societal experiments from 23 countries around the world that demonstrate a robust link between the prevalence of rule violations and intrinsic honesty. We developed an index of the 'prevalence of rule violations' (PRV) based on country-level data from the year 2003 of corruption, tax evasion and fraudulent politics. We measured intrinsic honesty in an anonymous die-rolling experiment[5]. We conducted the experiments with 2,568 young participants (students) who, due to their young age in 2003, could not have influenced PRV in 2003. We find individual intrinsic honesty is stronger in the subject pools of low PRV countries than those of high PRV countries. The details of lying patterns support psychological theories of honesty[6,7]. The results are consistent with theories of the cultural co-evolution of institutions and values[8], and show that weak institutions and cultural legacies[9–11] that generate rule violations not only have direct adverse economic consequences, but might also impair individual intrinsic honesty that is crucial for the smooth functioning of society.**

Good institutions that limit cheating and rule violations, such as corruption, tax evasion and political fraud are crucial for prosperity and development[12,13]. Yet, even very strong institutions cannot control all situations that may allow for cheating. Well-functioning societies also require the intrinsic honesty of citizens. Cultural characteristics, such as whether people see themselves as independent or part of a larger collective, that is, how individualist or collectivist[9] a society is, might also influence the prevalence of rule violations due to differences in the perceived scope of moral responsibilities, which is larger in more individualist cultures[10,14]. Here, we investigate how the prevalence of rule violations in a society and individual intrinsic honesty are linked. A variety of psychological, sociological and economic theories suggest causal pathways of how widespread practices of violating rules can affect individual honesty and the intrinsic willingness to follow rules.

Generally, processes of conformist transmission of values, beliefs and experiences influence individuals strongly and thereby can produce differences between social groups[15]. The extent to which people follow norms also depends on how prevalent norm violations are[3]. If cheating is pervasive in society and goes often unpunished, then people might view dishonesty in certain everyday affairs as justifiable without jeopardising their self-concept of being honest[2]. Experiencing frequent unfairness, an inevitable by-product of cheating, can also increase dishonesty[16]. Economic systems, institutions and business cultures shape people's ethical values[8,17,18], and can likewise impact individual honesty[19,20].

Ethical values, including honesty, are transmitted from prestigious people, peers and parents. People often take high-status individuals such as business leaders and celebrities as role models[21], and their cheating can set bad examples for dishonest practices[19]. Similarly, if politicians set bad examples by using fraudulent tactics like rigging elections, nepotism and embezzlement, then the honesty of citizens might suffer, because corruption is fostered in wider parts of society[13]. If many people work in the shadow economy and thereby evade taxes, peer effects might make cheating more acceptable[22]. If corruption is endemic in society, parents may recommend a positive attitude towards corruption and other acts of dishonesty and rule violations as a way to succeed in such an environment[4,23].

To measure the extent of society-wide practices of rule violations we constructed the PRV index. We focused on three broad types of rule violations: political fraud, tax evasion and corruption. We constructed PRV by calculating the principal component of three widely used country-level variables that all rest on comprehensive, often representative data sources to capture the important dimensions of the prevalence of rule violations that we are interested in: an indicator of political rights by Freedom House that measures the democratic quality of a country's political practices; the size of a country's shadow economy as a proxy for tax evasion; and corruption as measured by the World Bank's Control of Corruption Index (Supplementary Methods).

We constructed PRV for the 159 countries for which data are available for all three variables, the earliest year being 2003. We used the 2003 data to maximize the distance between the measurement of PRV and the point in time we ran the experiments (between 2011 and 2015, that is, at least 8 years after 2003). We use the 2003 data to ensure that our experimental participants could not have affected PRV in 2003 because they were still children at that time and therefore had been in no position to commit rule violations that influenced PRV in 2003. PRV in 2003 has a mean of 0 (s.d. = 1.46), and it ranges from −3.1 to 2.8 (higher values indicate higher prevalence of rule violations).

Our strategy was to conduct comparable experiments in 23 diverse countries with a distribution of PRV that resembles the world distribution of PRV. In the countries of our sample, PRV in 2003 ranges from −3.1 to 2.0, with a mean of −0.7 (s.d. = 1.52). Thus, the distribution of PRV in our sample is approximately representative of the world distribution of PRV with a slight bias towards lower PRV countries. The countries of our sample also vary strongly according to frequently used cultural indicators such as individualism and value orientations (Extended Data Table 1 and Supplementary Methods).

Our participants, all nationals of the respective country, were young people with comparable sociodemographic characteristics (students; mean age of 21.7 (s.d. = 3.3) years; 48% females; Supplementary Methods) who, due to their youth, had limited chances of being involved in political fraud, tax evasion or corruption, but might have been exposed to (or socialized into) certain attitudes towards (dis-) respecting rules[24].

[1]University of Nottingham, University Park, Nottingham NG7 2RD, UK. [2]CESifo, Schackstrasse 4, 80539 Munich, Germany. [3]IZA, Schaumburg-Lippe-Strasse 5-9, 53113 Bonn, Germany. [4]Yale University, 1 Prospect Street, New Haven, Connecticut 06510, USA.

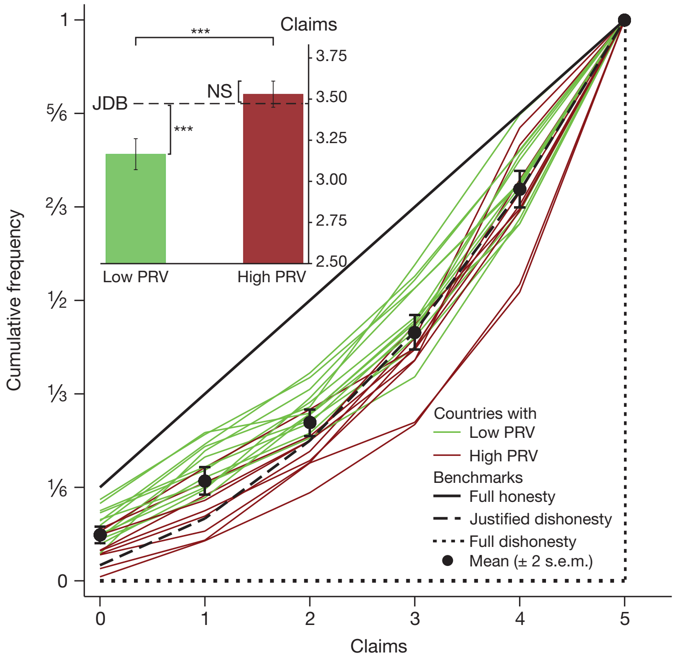


**Figure 1 | Distributions of reported die rolls.** Depicted are the cumulative distribution functions (CDFs) of amounts claimed compared to the CDFs of the full honesty, justified dishonesty and full dishonesty benchmarks. Green CDFs represent subject pools ($n_{\text{low}} = 14$) from countries with a below-average prevalence of rule violations (PRV; mean $\text{PRV}_{\text{low}} = -1.69$), and red CDFs represent subject pools ($n_{\text{high}} = 9$) from countries with above-average PRV (mean $\text{PRV}_{\text{high}} = 0.78$) out of 159 countries. Inset, the average claim (± 2 s.e.m.) is shown for subjects from below average ('low', $n_{\text{low}} = 1{,}211$) and above average ('high', $n_{\text{high}} = 1{,}357$) PRV countries. ***$P < 0.01$, two-sided $t$-tests; not significant (NS), $P > 0.14$. JDB, dishonesty benchmark.

Our experimental tool to measure intrinsic honesty was the 'die-in-a-cup' task[5]. Participants sat in a cubicle and were asked to roll a six-sided die placed in an opaque cup twice, but to report the first roll only. Die rolling was unobservable by anyone except the subject (Extended Data Fig. 1). Participants were paid according to the number they reported. Reporting a one earned the participant one money unit, claiming a two earned two money units, and so on, except that reporting a six earned nothing. Participants understood that reports were unverifiable. Across countries, money units reflected local purchasing power (Supplementary Methods). Thus, incentives in the experiment are the same for everyone, whether they live in a high or low PRV environment.

Although individual dishonesty is not detectable, aggregate behaviour is informative. In an honest subject pool, all numbers occur with a probability of one-sixth and the average claim is 2.5 money units. We refer to this as the 'full honesty' benchmark. By contrast, in the 'full dishonesty' benchmark, subjects follow their material incentives and claim 5 money units.

The die-in-a-cup task requires only a simple non-strategic decision, and it allows for gradual dishonesty predicted by psychological theories of honesty[6,7]. An experimentally tested theory of 'justified ethicality'[7] applied to our setting argues that many people have a desire to maintain an honest self-image. Lying about a die roll jeopardizes this self-image, but bending rules might not. Bending the rules is to report the higher of the two rolls, rather than the first roll as required. Reporting the better of two rolls implies the 'justified dishonesty' benchmark: claims of 0 should occur in $1/36 \approx 2.8\%$ of the cases (after rolling (6, 6)); claims of 1 should occur in $3/36 \approx 8.3\%$ (after (6,1) or (1,6) or (1,1)); claims of 2, 3, 4 and 5 should occur in 13.9%, 19.4%, 25% and 30.6% of cases, respectively.

Figure 1 illustrates the benchmarks, presented as cumulative distribution functions (CDFs). Figure 1 also shows the empirical CDF for each subject pool. CDFs are far away from full dishonesty. CDFs are also bent away from full honesty and cluster around the justified dishonesty benchmark. One-sample Kolmogorov–Smirnov tests for discrete data reject the null hypotheses of equality of CDFs with the full honesty benchmarks for every subject pool, but cannot reject the null hypothesis in 13 subject pools in comparisons with the justified dishonesty benchmark (Extended Data Fig. 2a).

Deviations from the justified dishonesty benchmark are related to PRV. The CDFs of subject pools from low PRV countries tend to be above the CDF implied by justified dishonesty, and also above those of most high PRV countries. Comparing the distributions of claims pooled for all low and high PRV countries, respectively, reveals a highly significant difference ($n_{\text{low}} = 1{,}211$, $n_{\text{high}} = 1{,}357$; $\chi^2(5) = 40.21$, $P < 0.001$). The pooled CDF from high PRV countries first-order stochastically dominates the pooled CDF from low PRV countries, that is, subjects from low PRV countries are more honest than subjects from high PRV countries. The pooled CDF from low PRV countries also lies significantly above justified dishonesty (Kolmogorov–Smirnov test, $d = 0.103$, $P < 0.001$), whereas the pooled CDF from high PRV countries tends to be slightly below it (Kolmogorov–Smirnov test, $d = 0.058$, $P < 0.001$; Extended Data Fig. 2b and Supplementary Information).

The inset in Fig. 1 illustrates the implications of these patterns in terms of average claims. Subjects from low PRV countries claim 3.17 money units (s.d. = 1.67), that is, 0.67 money units more than under full honesty. Subjects from high PRV countries claim 3.53 money units (s.d. = 1.49) or 1.03 money units more than under full honesty. This difference in claims is significant ($t$-test, $t = 5.84$, two-sided $P < 0.001$); it also holds at the country level ($n = 23$; Mann–Whitney test, $z = 3.40$, two-sided $P < 0.001$). Justified dishonesty implies an expected claim of 3.47 money units. The average claim in high PRV countries is not significantly different from this benchmark (one-sample $t$-test, $n_{\text{high}} = 1{,}357$, $t = 1.48$, two-sided $P = 0.140$), but is significantly lower in low PRV countries (one-sample $t$-test, $n_{\text{low}} = 1{,}211$, $t = 6.35$, two-sided $P < 0.001$).

Next we looked at four measures of dishonesty that can be derived from our task (Supplementary Information) and related them to country-level PRV (Fig. 2). A first measure of dishonesty is mean claim, which ranges from 2.96 money units to 3.96 money units across countries (mean = 3.32 money units, s.d. = 0.26; Kruskal–Wallis test, $\chi^2(22) = 75.2$, $P < 0.001$). PRV and mean claim are strongly positively related (Fig. 2a).

A second measure is the frequency of high claims for reported numbers 3, 4 and 5, which should occur at 50% if people are honest and at 75% under justified dishonesty. Frequencies range from 61.0% to 84.3% (mean = 71.8%, s.d. = 5.7%; $\chi^2(22) = 45.0$, $P = 0.003$). PRV and high claims are strongly positively associated (Fig. 2b).

The incentive is to claim 5, irrespective of the number actually rolled. Thus, the fraction of income maximizers provides our third measure of dishonesty. It is estimated from the fraction of people who reported 5 (highest claim) minus the expected rate of actual rolls of 5 (16.7%). To account for income maximizers who actually rolled a 5, the difference has to be multiplied by 6/5 (ref. 5). The rate of income maximizers ranges from 0.3% to 38.3% across subject pools (mean = 16.2%, s.d. = 9.4%; $\chi^2(22) = 72.4$, $P < 0.001$). Given that PRV captures rule violations for selfish gains and evidence suggesting rule breakers tend to be more selfish[25], we predict that income maximizers is positively correlated with PRV. We find, however, that they are unrelated (Fig. 2c). Thus, a society's PRV does not systematically affect maximal cheating in this experiment.

This result is in contrast to the observation that the estimated fraction of fully honest people and PRV are significantly negatively related (Fig. 2d). The fraction of fully honest people, our fourth measure, is estimated from 'no claim', that is, reports of rolling 6. A report of 6 is most likely honest and honest reports can occur for all numbers. Therefore, the fraction of fully honest people can be estimated as the fraction of people reporting 6 multiplied by six. Across subject

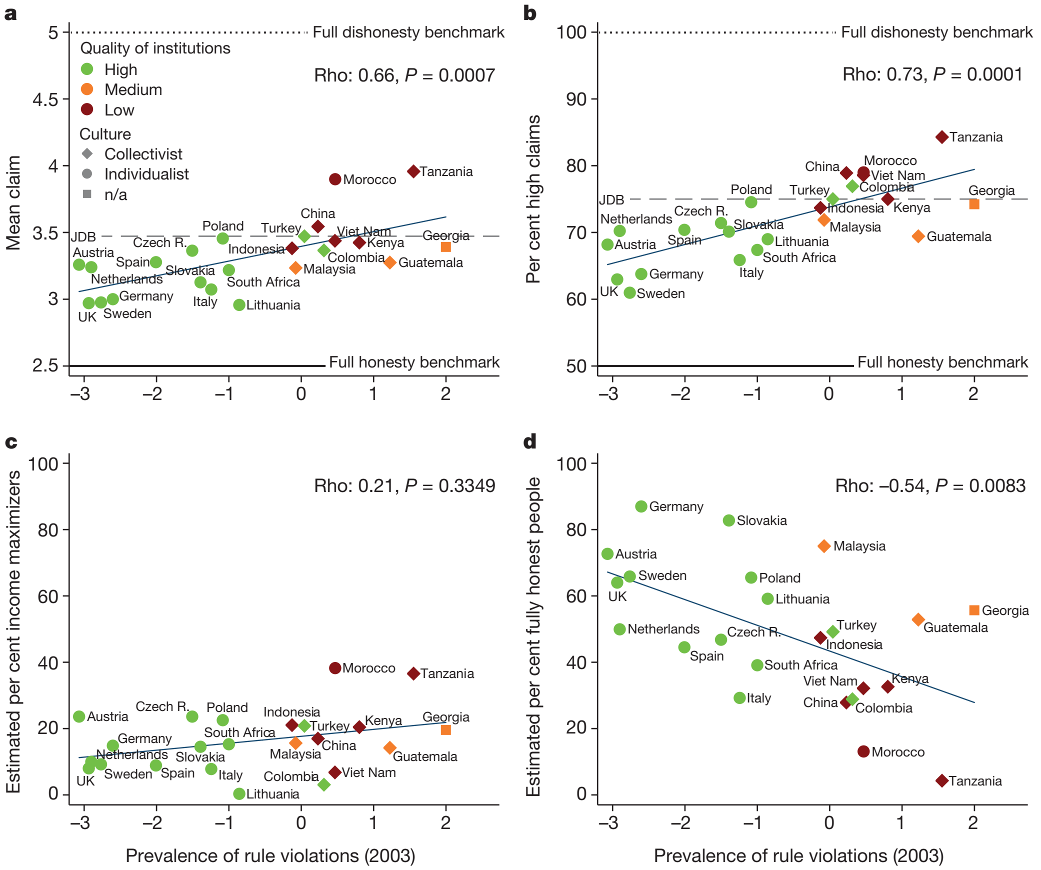


**Figure 2 | Measures of honesty and the prevalence of rule violations in society.** Shown are scatter plots of four measures of honesty and PRV at country level ($n = 23$); higher values indicate more rule violations. **a**, Mean claim. **b**, Per cent high claims of 3, 4 and 5 money units. **c**, Per cent income maximizers estimated from the fraction of people claiming 5 money units. **d**, Per cent fully honest people estimated from the fraction of people claiming 0 money units. Rho is the Spearman rank correlation based on country means. JDB is the justified dishonesty benchmark (not defined for **c** and **d**). Colour coding refers to the quality of institutions as measured by Constraints on Executives; shapes distinguish between countries classified as collectivist or individualist. PRV is negatively correlated with Constraints on Executives and Individualism (Supplementary Information); this also holds in our sample (Constraint on Executive: rho = −0.76, $n = 23$, $P < 0.0001$; Individualism: rho = −0.79, $n = 22$, $P < 0.0001$).

pools, fully honest people range from 4.3% to 87% (mean = 48.9%, s.d. = 21.3%; $\chi^2(22) = 42.1$, $P = 0.006$). In societies with high levels of PRV, fewer people are fully honest than in societies with low levels of PRV.

Regression analyses that control for individual attitudes to honesty and beliefs in the fairness of others, as well as for sociodemographics confirm the robustness of our results (Extended Data Table 2 and Supplementary Information). Sociodemographic variables, including gender, are generally insignificant. Stronger individual norms of honesty significantly reduce mean claim, high claim and highest claim. Beliefs in the fairness of others only significantly reduce highest claim.

Results are also robust using the earliest available data related to PRV, corruption in 1996; using 'Government Effectiveness', a proxy for bureaucratic quality and material security[11] and measures of institutional quality that emphasize law enforcement (rules) and not actual compliance and which extend far into the past, so they are most likely not influenced even by parents (Extended Data Fig. 3a–d and Supplementary Information).

Given that the experiment holds the rules and incentives constant for everyone, the large differences across subject pools are also consistent with a cultural transmission of norms of honesty and rule following through the generations[4,15,23] and a co-evolution of norms and institutions[8]. Societies with higher material security, as measured by Government Effectiveness, tend to be more individualist[11], and more individualist societies tend to have less corruption[10]. Consistent with this, we find that subject pools from individualist societies have lower claims than subject pools from more collectivist societies and also from more traditional societies and societies with survival-related values (Extended Data Fig. 4a–c and Supplementary Information). Further econometric analyses developed in economic literature on culture and institutions[14] applied to PRV support the argument that both the quality of institutions, as well as culture (individualism) are highly significantly (and likely causally) correlated with PRV (Extended Data Table 3 and Supplementary Information).

Taken together, our results suggest that institutions and cultural values influence PRV, which, through various theoretically predicted and experimentally tested pathways[2,11,16,19,20,22–26], impact on people's intrinsic honesty and rule following. Our experiments from around the globe also provide support for arguments that for many people lying is psychologically costly[27–30]. More specifically, theories of honesty posit that many people are either honest, or (self-deceptively[1]) bend rules or lie gradually to an extent that is compatible with maintaining an honest self-image[6,7]. Evidence for lying aversion and honest self-concepts has been mostly confined to western societies with low PRV values[30]. Our expanded scope of societies therefore provides important support and qualifications for the generalizability of these theories—people benchmark their justifiable dishonesty with the extent of dishonesty they see in their societal environment.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 9 April 2015; accepted 25 January 2016.**
**Published online 9 March 2016.**

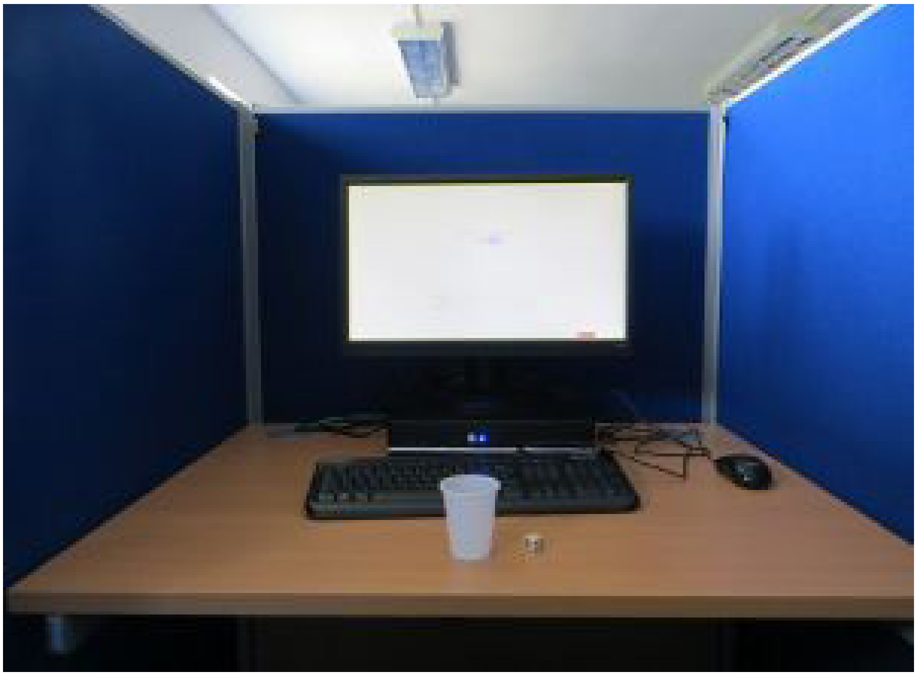
1. Trivers, R. L. *Deceit and Self-deception: Fooling Yourself the Better to Fool Others* (Penguin, 2014).
2. Gino, F., Ayal, S. & Ariely, D. Contagion and differentiation in unethical behavior: the effect of one bad apple on the barrel. *Psychol. Sci.* **20,** 393–398 (2009).
3. Keizer, K., Lindenberg, S. & Steg, L. The spreading of disorder. *Science* **322,** 1681–1685 (2008).
4. Hauk, E. & Saez-Marti, M. On the cultural transmission of corruption. *J. Econ. Theory* **107,** 311–335 (2002).
5. Fischbacher, U. & Föllmi-Heusi, F. Lies in disguise—an experimental study on cheating. *J. Eur. Econ. Assoc.* **11,** 525–547 (2013).
6. Mazar, N., Amir, O. & Ariely, D. The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res.* **45,** 633–644 (2008).
7. Shalvi, S., Dana, J., Handgraaf, M. J. J. & De Dreu, C. K. W. Justified ethicality: observing desired counterfactuals modifies ethical perceptions and behavior. *Organ. Behav. Hum. Decis. Process.* **115,** 181–190 (2011).
8. Bowles, S. Is liberal society a parasite on tradition? *Philos. Public Aff.* **39,** 46–81 (2011).
9. Greif, A. Cultural beliefs and the organization of society: a historical reflection on collectivist and individualist societies. *J. Polit. Econ.* **102,** 912–950 (1994).
10. Mazar, N. & Aggarwal, P. Greasing the palm: can collectivism promote bribery? *Psychol. Sci.* **22,** 843–848 (2011).
11. Hruschka, D. *et al.* Impartial institutions, pathogen stress and the expanding social network. *Hum. Nat.* **25,** 567–579 (2014).
12. Besley, T. & Persson, T. *Pillars of Prosperity: the Political Economics of Development Clusters* (Princeton Univ. Press, 2011).
13. Heywood, P. M. *Routledge Handbook of Political Corruption* (Routledge, 2014).
14. Tabellini, G. Institutions and culture. *J. Eur. Econ. Assoc.* **6,** 255–294 (2008).
15. Henrich, J. & Boyd, R. The evolution of conformist transmission and the emergence of between-group differences. *Evol. Hum. Behav.* **19,** 215–241 (1998).
16. Houser, D., Vetter, S. & Winter, J. Fairness and cheating. *Eur. Econ. Rev.* **56,** 1645–1655 (2012).
17. Gintis, H. & Khurana, R. in *Moral Markets: the Critical Role of Values in the Economy* (ed. Zak, P. J.) (Princeton Univ. Press, 2008).
18. Crittenden, V., Hanna, R. & Peterson, R. Business students' attitudes toward unethical behavior: a multi-country comparison. *Mark. Lett.* **20,** 1–14 (2009).
19. Cohn, A., Fehr, E. & Marechal, M. A. Business culture and dishonesty in the banking industry. *Nature* **516,** 86–89 (2014).
20. Weisel, O. & Shalvi, S. The collaborative roots of corruption. *Proc. Natl Acad. Sci. USA* **112,** 10651–10656 (2015).
21. Henrich, J. & Gil-White, F. J. The evolution of prestige: freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evol. Hum. Behav.* **22,** 165–196 (2001).
22. Lefebvre, M., Pestieau, P., Riedl, A. & Villeval, M. Tax evasion and social information: an experiment in Belgium, France, and The Netherlands. *Int. Tax Public Finance* **22,** 401–425 (2015).
23. Tabellini, G. The scope of cooperation: values and incentives. *Q. J. Econ.* **123,** 905–950 (2008).
24. Barr, A. & Serra, D. Corruption and culture: an experimental analysis. *J. Public Econ.* **94,** 862–869 (2010).
25. Kimbrough, E. O. & Vostroknutov, A. Norms make preferences social. *J. Eur. Econ. Assoc.* http://dx.doi.org/10.1111/jeea.12152 (2015).
26. Peysakhovich, A. & Rand, D. G. Habits of virtue: creating norms of cooperation and defection in the laboratory. *Manage. Sci.* http://dx.doi.org/10.1287/mnsc.2015.2168 (2015).
27. Gneezy, U. Deception: the role of consequences. *Am. Econ. Rev.* **95,** 384–394 (2005).
28. Abeler, J., Becker, A. & Falk, A. Representative evidence on lying costs. *J. Public Econ.* **113,** 96–104 (2014).
29. Pascual-Ezama, D. *et al.* Context-dependent cheating: experimental evidence from 16 countries. *J. Econ. Behav. Organ.* **116,** 379–386 (2015).
30. Rosenbaum, S. M., Billinger, S. & Stieglitz, N. Let's be honest: a review of experimental evidence of honesty and truth-telling. *J. Econ. Psychol.* **45,** 181–196 (2014).

**Supplementary Information** is available in the online version of the paper.

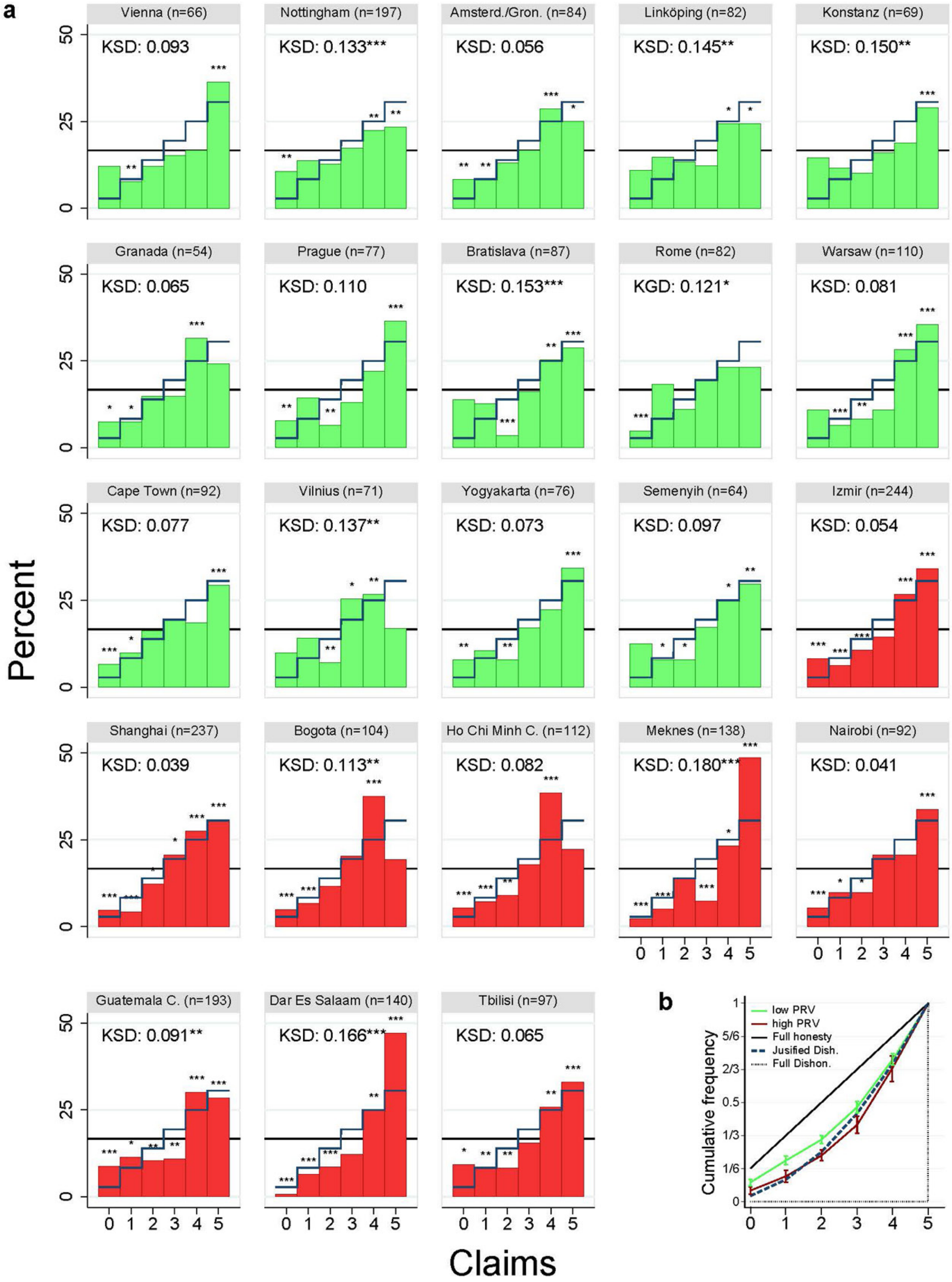
**Acknowledgements** We thank A. Arechar, A. Barr, B. Beranek, M. Eberhardt, E. von Essen, E. Fehr, U. Fischbacher, M. García-Vega, A. Greif, B. Herrmann, F. Kölle, L. Molleman, D. Rand, K. Schmelz, S. Shalvi, P. Thiemann, C. Thöni, O. Weisel and seminar audiences for comments. Support under ERC-AdG 295707 COOPERATION and the ESRC Network on Integrated Behavioural Science (NIBS, ES/K002201/1) is gratefully acknowledged. We thank numerous helpers (Supplementary Information) for their support in implementing the experiments.

**Author Contributions** S.G. and J.S. developed the research ideas and designed the study; J.S. conducted the experiment and analysed data. S.G. and J.S. wrote the manuscript.

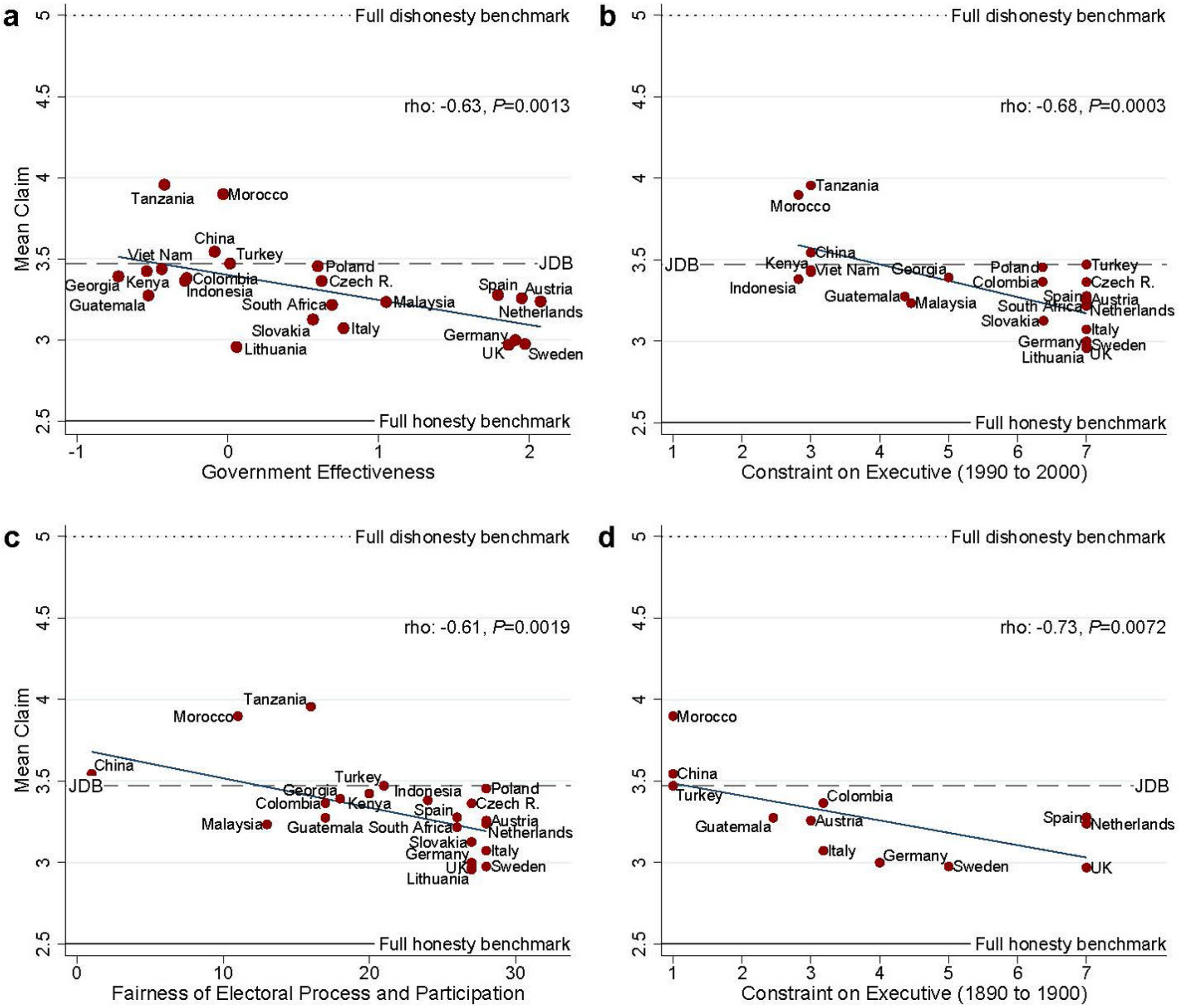
**Author Information** The data and code for the statistical analyses are stored in Dryad Data package title Intrinsic Honesty across Societies, http://dx.doi.org/10.5061/dryad.9k358. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.G. (simon.gaechter@nottingham.ac.uk) or J.S. (jonathan.schulz@yale.edu).



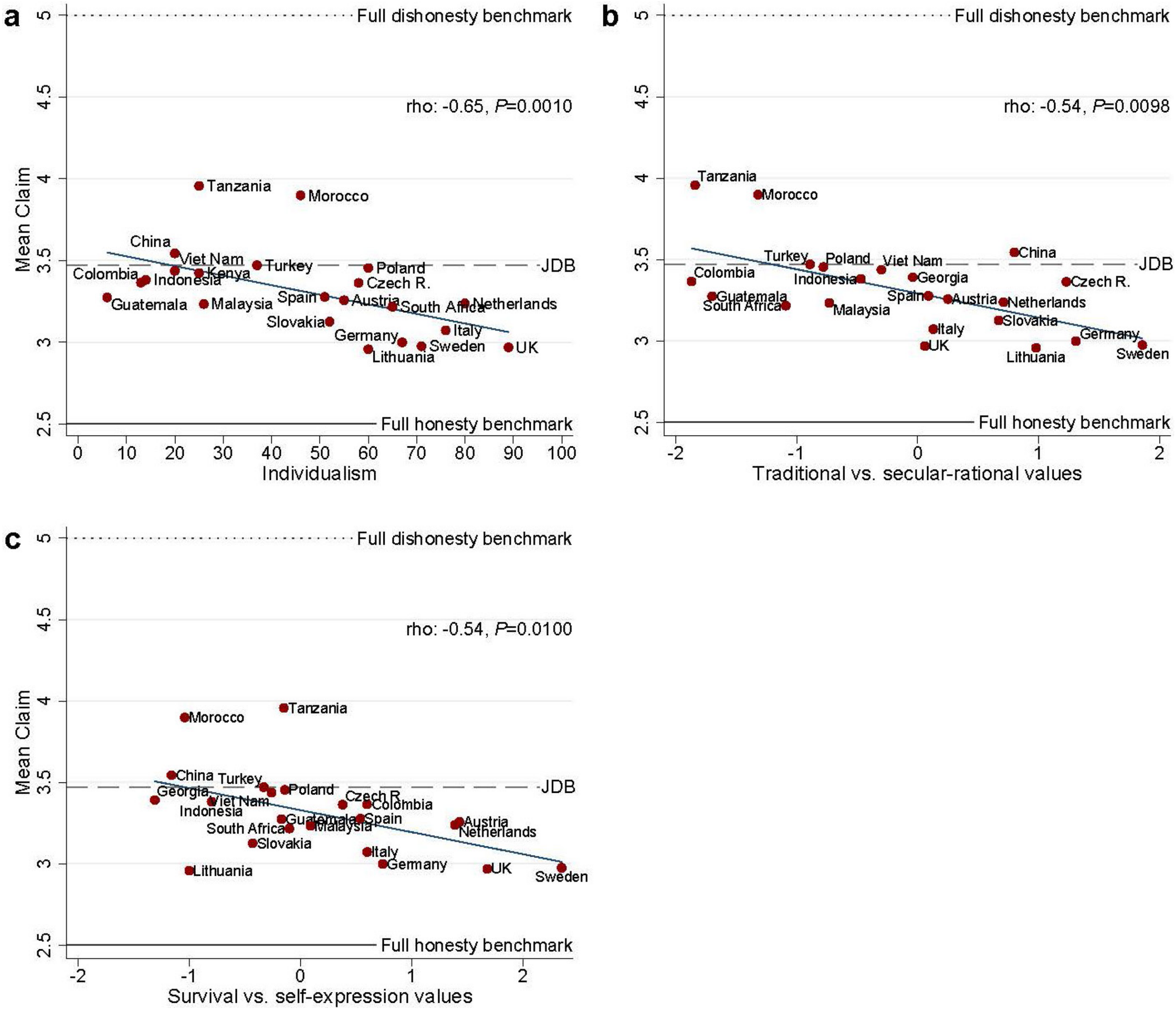
**Extended Data Figure 1 | The die-in-a-cup task.** Experiment following Fischbacher and Föllmi-Heusi[5]. Participants ($n = 2{,}568$ from 23 countries) were asked to roll the die twice in the cup and to report the first roll. Payment is according to reported roll, except that reporting a 6 earns 0 money units (across subject pools, money units in local currency are adjusted to equalize purchasing power). We used the same set of dice in all subject pools, and we also tested the dice for bias. The procedures followed established rules in cross-cultural experimental economics. See Supplementary Information for further details. This picture was taken by J.S. in the experimental laboratory of the University of Nottingham.



**Extended Data Figure 2 | Distribution of claims. a**, Distribution per subject pool. Subject pools are ordered by country PRV. The first 14 subject pools (in green) are from 'low' (below-average) PRV countries; the last 9 subject pools (in red) are from 'high' (above-average) PRV countries relative to the world sample of 159 countries. The horizontal line refers to the uniform distribution implied by honest reporting and the step function to the distribution implied by the justified dishonesty benchmark (JDB). For each subject pool, we report the one-sample Kolmogorov–Smirnov test (KS) for discrete data in comparison with JDB (KSD is the KS *d* value). Asterisks above bars refer to binomial tests comparing the frequency of a particular claim with its predicted value under a uniform distribution. **b**, Cumulative distributions for pooled data from subject pools from low and high PRV countries, respectively. See Supplementary Information for further information. **P* < 0.1, ***P* < 0.05, ****P* < 0.01.

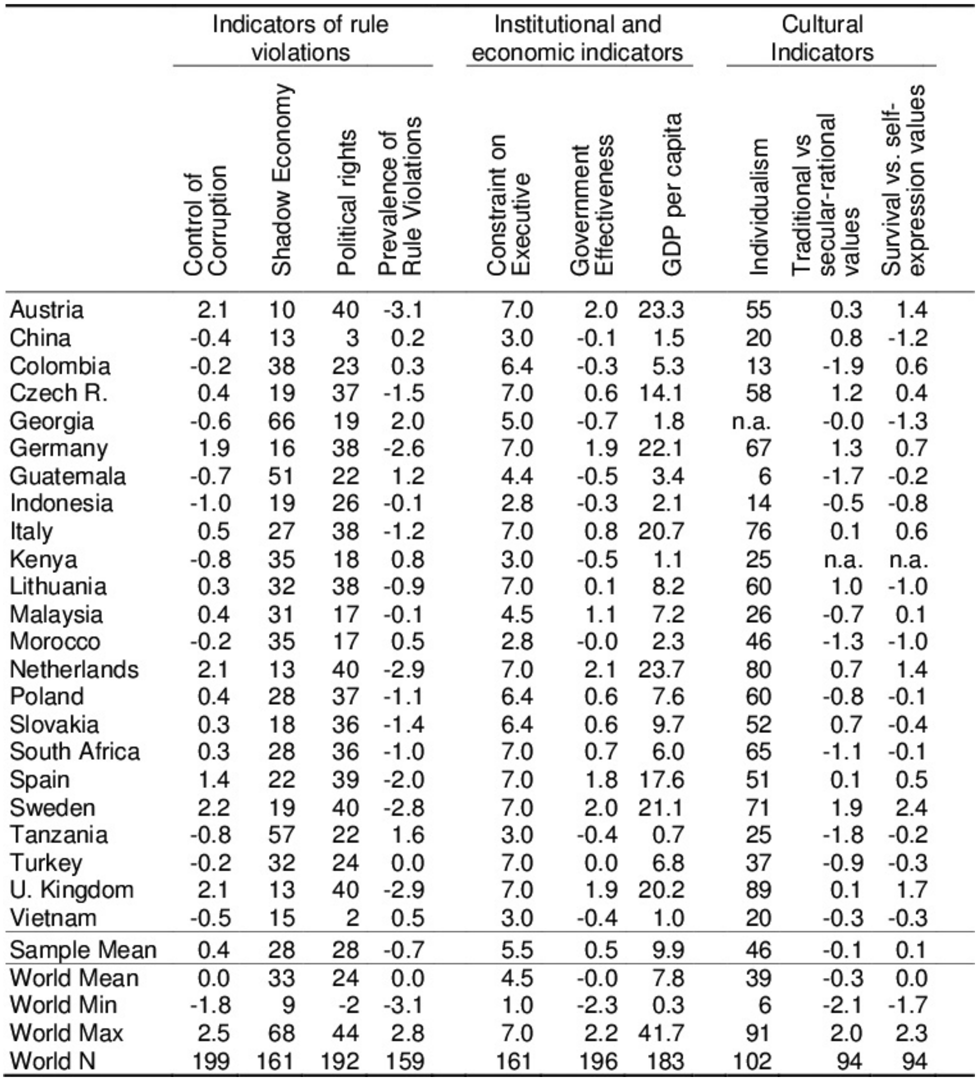


**Extended Data Figure 3 | Association between indicators of institutional quality and intrinsic honesty as measured by mean claim.** The solid line is a linear fit. The JDB is indicated by a dashed line. Rho indicates Spearman rank order correlation coefficients. **a**–**d**, Mean claim is negatively related to Government Effectiveness (**a**), Constraint on executive (**b**), ‘fairness of electoral process and participation’ (**c**) and Constraint on Executive (**d**) using the averages of the years 1890 to 1900 as a measure for distant institutional quality. See Extended Data Table 1 and Supplementary Information for data description, references and further analyses.



**Extended Data Figure 4 | Association between cultural indicators and intrinsic honesty as measured by mean claim.** The solid line is a linear fit. The JDB is indicated by a dashed line. Rho indicates Spearman rank order correlation coefficients. **a**–**c**, Mean claim is negatively related to individualism (**a**), traditional versus secular-rational values (**b**), and survival versus self-expression values (**c**). See Extended Data Table 1 and Supplementary Information for data description, references and further analyses.

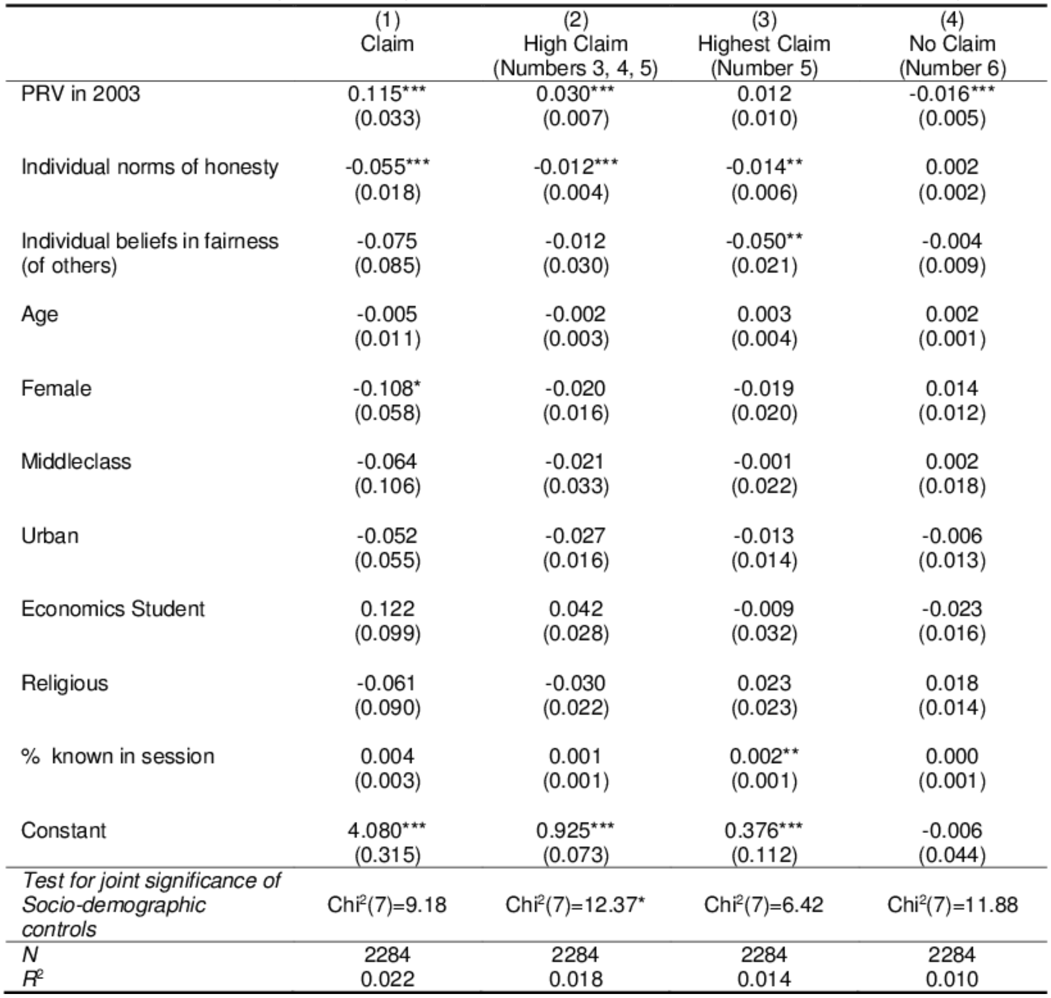
**Extended Data Table 1 | Measures of prevalence of rule violations, economic and institutional variables, as well as cultural background of our subject pools**



| | Indicators of rule violations | | | | Institutional and economic indicators | | | Cultural Indicators | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Control of Corruption | Shadow Economy | Political rights | Prevalence of Rule Violations | Constraint on Executive | Government Effectiveness | GDP per capita | Individualism | Traditional vs secular-rational values | Survival vs. self-expression values |
| Austria | 2.1 | 10 | 40 | -3.1 | 7.0 | 2.0 | 23.3 | 55 | 0.3 | 1.4 |
| China | -0.4 | 13 | 3 | 0.2 | 3.0 | -0.1 | 1.5 | 20 | 0.8 | -1.2 |
| Colombia | -0.2 | 38 | 23 | 0.3 | 6.4 | -0.3 | 5.3 | 13 | -1.9 | 0.6 |
| Czech R. | 0.4 | 19 | 37 | -1.5 | 7.0 | 0.6 | 14.1 | 58 | 1.2 | 0.4 |
| Georgia | -0.6 | 66 | 19 | 2.0 | 5.0 | -0.7 | 1.8 | n.a. | -0.0 | -1.3 |
| Germany | 1.9 | 16 | 38 | -2.6 | 7.0 | 1.9 | 22.1 | 67 | 1.3 | 0.7 |
| Guatemala | -0.7 | 51 | 22 | 1.2 | 4.4 | -0.5 | 3.4 | 6 | -1.7 | -0.2 |
| Indonesia | -1.0 | 19 | 26 | -0.1 | 2.8 | -0.3 | 2.1 | 14 | -0.5 | -0.8 |
| Italy | 0.5 | 27 | 38 | -1.2 | 7.0 | 0.8 | 20.7 | 76 | 0.1 | 0.6 |
| Kenya | -0.8 | 35 | 18 | 0.8 | 3.0 | -0.5 | 1.1 | 25 | n.a. | n.a. |
| Lithuania | 0.3 | 32 | 38 | -0.9 | 7.0 | 0.1 | 8.2 | 60 | 1.0 | -1.0 |
| Malaysia | 0.4 | 31 | 17 | -0.1 | 4.5 | 1.1 | 7.2 | 26 | -0.7 | 0.1 |
| Morocco | -0.2 | 35 | 17 | 0.5 | 2.8 | -0.0 | 2.3 | 46 | -1.3 | -1.0 |
| Netherlands | 2.1 | 13 | 40 | -2.9 | 7.0 | 2.1 | 23.7 | 80 | 0.7 | 1.4 |
| Poland | 0.4 | 28 | 37 | -1.1 | 6.4 | 0.6 | 7.6 | 60 | -0.8 | -0.1 |
| Slovakia | 0.3 | 18 | 36 | -1.4 | 6.4 | 0.6 | 9.7 | 52 | 0.7 | -0.4 |
| South Africa | 0.3 | 28 | 36 | -1.0 | 7.0 | 0.7 | 6.0 | 65 | -1.1 | -0.1 |
| Spain | 1.4 | 22 | 39 | -2.0 | 7.0 | 1.8 | 17.6 | 51 | 0.1 | 0.5 |
| Sweden | 2.2 | 19 | 40 | -2.8 | 7.0 | 2.0 | 21.1 | 71 | 1.9 | 2.4 |
| Tanzania | -0.8 | 57 | 22 | 1.6 | 3.0 | -0.4 | 0.7 | 25 | -1.8 | -0.2 |
| Turkey | -0.2 | 32 | 24 | 0.0 | 7.0 | 0.0 | 6.8 | 37 | -0.9 | -0.3 |
| U. Kingdom | 2.1 | 13 | 40 | -2.9 | 7.0 | 1.9 | 20.2 | 89 | 0.1 | 1.7 |
| Vietnam | -0.5 | 15 | 2 | 0.5 | 3.0 | -0.4 | 1.0 | 20 | -0.3 | -0.3 |
| Sample Mean | 0.4 | 28 | 28 | -0.7 | 5.5 | 0.5 | 9.9 | 46 | -0.1 | 0.1 |
| World Mean | 0.0 | 33 | 24 | 0.0 | 4.5 | -0.0 | 7.8 | 39 | -0.3 | 0.0 |
| World Min | -1.8 | 9 | -2 | -3.1 | 1.0 | -2.3 | 0.3 | 6 | -2.1 | -1.7 |
| World Max | 2.5 | 68 | 44 | 2.8 | 7.0 | 2.2 | 41.7 | 91 | 2.0 | 2.3 |
| World N | 199 | 161 | 192 | 159 | 161 | 196 | 183 | 102 | 94 | 94 |

Data are country-level averages. Detailed descriptions, data sources and references are in the Supplementary Information. Control of corruption is a standard measure of corruption; higher values indicate more corruption. Shadow economy is measured in percent of the size of a country's gross domestic product (GDP). Political rights measures the fairness of electoral processes, political pluralism and participation, and the functioning of government; higher scores indicate higher level of political rights. Prevalence of rule violations is our self-constructed indicator based on a principal component analysis of control of corruption, shadow economy and political rights. Government Effectiveness measures the quality of public service, independence from political pressure and policy implementation; higher values indicate higher effectiveness. Constraint on Executive measures the institutionalised limitations on the arbitrary use of power by the executive; higher values indicate better control. GDP per capita (average of 1990 to 2000) is measured in units of US dollars $1,000 (purchasing power parity (PPP)). Individualism measures how important the individual is relative to the collective; higher values indicate higher individualism. Traditional versus secular-rational values measures the importance of values such as respect for authorities; higher scores indicate more secular values. Survival versus self-expression values measure the importance of values surrounding physical and economic security; lower scores indicate survival values are relatively more important than self-expression values. World mean and sample mean are the respective averages of country means.

**Extended Data Table 2 | Regression analysis of societal and individual determinants of dishonesty**



| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | Claim | High Claim | Highest Claim | No Claim |
| | | (Numbers 3, 4, 5) | (Number 5) | (Number 6) |
| PRV in 2003 | 0.115*** | 0.030*** | 0.012 | -0.016*** |
| | (0.033) | (0.007) | (0.010) | (0.005) |
| Individual norms of honesty | -0.055*** | -0.012*** | -0.014** | 0.002 |
| | (0.018) | (0.004) | (0.006) | (0.002) |
| Individual beliefs in fairness (of others) | -0.075 | -0.012 | -0.050** | -0.004 |
| | (0.085) | (0.030) | (0.021) | (0.009) |
| Age | -0.005 | -0.002 | 0.003 | 0.002 |
| | (0.011) | (0.003) | (0.004) | (0.001) |
| Female | -0.108* | -0.020 | -0.019 | 0.014 |
| | (0.058) | (0.016) | (0.020) | (0.012) |
| Middleclass | -0.064 | -0.021 | -0.001 | 0.002 |
| | (0.106) | (0.033) | (0.022) | (0.018) |
| Urban | -0.052 | -0.027 | -0.013 | -0.006 |
| | (0.055) | (0.016) | (0.014) | (0.013) |
| Economics Student | 0.122 | 0.042 | -0.009 | -0.023 |
| | (0.099) | (0.028) | (0.032) | (0.016) |
| Religious | -0.061 | -0.030 | 0.023 | 0.018 |
| | (0.090) | (0.022) | (0.023) | (0.014) |
| % known in session | 0.004 | 0.001 | 0.002** | 0.000 |
| | (0.003) | (0.001) | (0.001) | (0.001) |
| Constant | 4.080*** | 0.925*** | 0.376*** | -0.006 |
| | (0.315) | (0.073) | (0.112) | (0.044) |
| *Test for joint significance of Socio-demographic controls* | $\text{Chi}^2(7)=9.18$ | $\text{Chi}^2(7)=12.37$* | $\text{Chi}^2(7)=6.42$ | $\text{Chi}^2(7)=11.88$ |
| *N* | 2284 | 2284 | 2284 | 2284 |
| $R^2$ | 0.022 | 0.018 | 0.014 | 0.010 |

The explanatory variables are the scores of a country's prevalence of rule violations in 2003; participants' individual norms of honesty (based on individual opinions about justifiableness of various acts of cheating; higher scores indicate stronger norms); participants' beliefs in fairness (the perceived fairness of most others; a higher score indicates a higher belief). Sociodemographic controls include age; dummies for sex, urban residency, middle class status, being an economics student, and being religious; and the percentage of other participants known to a participant. Detailed data description and rationale are in the Supplementary Methods. Chi-square tests reveal that sociodemographic controls are jointly insignificant in all models except model 2, where they are weakly significant. The estimation method is ordinary least squares (OLS) with bootstrapped standard errors clustered on countries. The results are robust to various specifications (Supplementary Information). **P* < 0.10, ***P* < 0.05, ****P* < 0.01.

**Extended Data Table 3 | Institutional and cultural determinants of PRV**

| | (1) | (2) | (3) | (4) | (5) | (6) | (7)<br>IV: Sett.<br>Mortality | (8)<br>IV: Gram.<br>Rule | (9)<br>IV: Gen.<br>Dist. | (10)<br>IV: Gen. Dist.<br>+<br>Gram. Rule |
|---|---|---|---|---|---|---|---|---|---|---|
| Const. on Executive<br>(1990 to 2000) | -0.25***<br>(0.05) | | -0.23***<br>(0.07) | -0.21***<br>(0.05) | -0.09***<br>(0.03) | -0.25***<br>(0.05) | **-0.72*****<br>**(0.12)** | -0.25**<br>(0.11) | -0.23***<br>(0.08) | -0.25**<br>(0.11) |
| Individualism | -0.03***<br>(0.00) | -0.03***<br>(0.01) | -0.02***<br>(0.01) | -0.02***<br>(0.00) | -0.01**<br>(0.00) | -0.03***<br>(0.00) | | **-0.06***<br>**(0.03)** | **-0.05****<br>**(0.03)** | **-0.06****<br>**(0.03)** |
| Const. on Executive<br>(1890 to 1900) | | -0.26***<br>(0.06) | | | | | | | | |
| Primary Education<br>(1930) | | | -0.02***<br>(0.00) | | | | | 0.01<br>(0.02) | 0.00<br>(0.02) | 0.01<br>(0.02) |
| GDP p. capita<br>(PPP in $ 1000) | | | | -0.07***<br>(0.01) | | | | | | |
| Gov. Effective-<br>ness (2000) | | | | | -1.10***<br>(0.07) | | | | | |
| Ethnolinguistic<br>Fractionalization | | | | | | 0.41<br>(0.38) | | | | |
| Constant | 2.14***<br>(0.26) | 1.67***<br>(0.17) | 2.20***<br>(0.30) | 2.02***<br>(0.22) | 0.59***<br>(0.19) | 1.91***<br>(0.33) | 3.79***<br>(0.53) | 2.69***<br>(0.56) | 2.67***<br>(0.51) | 2.68***<br>(0.51) |
| *Controls for Legal Origin* | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| *N* | 96 | 44 | 79 | 96 | 96 | 96 | 60 | 59 | 79 | 59 |
| *R*² | 0.681 | 0.810 | 0.785 | 0.824 | 0.904 | 0.685 | 0.131 | 0.633 | 0.673 | 0.652 |
| *1st-stage F-stat* | | | | | | | 12.4*** | 60.3*** | 51.7*** | 68.4*** |
| *Overid test p-value* | | | | | | | | | | 0.907 |

Dependent variable is PRV in 2003. Our approach follows recent advances in the economic literature on institutions and culture (see Supplementary Information for details and references). Models 1 to 6 are OLS; models 7 to 10 use instrumental variables to identify causal relations. All regressions control for legal origin (French, British, German, Scandinavian). Model 1 shows that both a frequently used measure for institutional quality (Constraint on Executive) and a frequently used measure for culture (Individualism) are significantly correlated with PRV. Model 2 shows that past institutional quality (Constraint on Executive in 1890–1900) can have long-lasting effects on PRV. Models 3 to 6 control for important variables proposed in the literature. Models 7 to 10 report the results from instrumental variable estimation (instrumented variables are in bold); the instruments are assumed to have no direct impact on PRV but only on the explanatory variable, and thereby allow identifying a causal effect of either institutions (as measured by Constraint on Executive) or culture (as measured by Individualism) on PRV. Model 7 instruments institution with 'settler mortality' in European colonies (1600–1875). To preserve degrees of freedom we do not include Individualism. Model 8 uses language (grammatical rules) and model 9 uses genetic distance as an instrument for culture. Model 10 uses both instruments. Models 7 to 10 suggest causal effects of both the quality of institutions and culture (Individualism) on PRV. *$P < 0.10$, **$P < 0.05$, ***$P < 0.01$.

doi:10.1038/nature16990

# Impact of meat and Lower Palaeolithic food processing techniques on chewing in humans

Katherine D. Zink[1] & Daniel E. Lieberman[1]

**The origins of the genus *Homo* are murky, but by *H. erectus*, bigger brains and bodies had evolved that, along with larger foraging ranges, would have increased the daily energetic requirements of hominins[1,2]. Yet *H. erectus* differs from earlier hominins in having relatively smaller teeth, reduced chewing muscles, weaker maximum bite force capabilities, and a relatively smaller gut[3–5]. This paradoxical combination of increased energy demands along with decreased masticatory and digestive capacities is hypothesized to have been made possible by adding meat to the diet[6–8], by mechanically processing food using stone tools[7,9,10], or by cooking[11,12]. Cooking, however, was apparently uncommon until 500,000 years ago[13,14], and the effects of carnivory and Palaeolithic processing techniques on mastication are unknown. Here we report experiments that tested how Lower Palaeolithic processing technologies affect chewing force production and efficacy in humans consuming meat and underground storage organs (USOs). We find that if meat comprised one-third of the diet, the number of chewing cycles per year would have declined by nearly 2 million (a 13% reduction) and total masticatory force required would have declined by 15%. Furthermore, by simply slicing meat and pounding USOs, hominins would have improved their ability to chew meat into smaller particles by 41%, reduced the number of chews per year by another 5%, and decreased masticatory force requirements by an additional 12%. Although cooking has important benefits, it appears that selection for smaller masticatory features in *Homo* would have been initially made possible by the combination of using stone tools and eating meat.**

Two derived human behaviours are meat eating and food processing. Archaeological and palaeontological evidence indicate that hominins began to increase meat consumption by at least 2.6 million years ago (Ma) (ref. 7), and until the invention of agriculture, meat was an indispensable component of human diets[15]. Archaeological data also indicate that hominins fabricated stone tools by 3.3 Ma (ref. 10), learned to control fire by 1 Ma (ref. 13), and started to cook on a regular basis by at least 0.5 Ma (refs 13, 14). Today, humans process most of their food in some way before ingestion. Yet, despite the importance of meat eating and food processing, little is currently known about the degree to which these novel behaviours altered selection on the hominin masticatory apparatus. Multiple lines of evidence indicate that the australopith ancestors of *Homo* consumed lots of mechanically demanding plant foods[16] and probably resembled great apes in spending a substantial proportion of the day feeding and chewing, approximately an order of magnitude more than non-industrial humans[17]. Maximum bite force capabilities in early *Homo* were less than half that of australopiths[3], and while *H. habilis* retained many primitive masticatory features, including large, thick post-canine teeth, *H. erectus* had considerably smaller post-canines, along with smaller faces. These derived masticatory features suggest that the genus *Homo* consumed foods that were easier to eat, requiring fewer, less forceful chews and reducing the need for high maximum bite forces. But it has been unclear to what extent these shifts were made possible by meat, by mechanical processing, or by cooking.

Efforts to understand how diets differed between australopiths and early *Homo* have focused on increased consumption of meat (muscle tissue) and the benefits of cooking[6,12,18]. Muscle tissue is calorically denser than most plant foods, but is difficult to chew with low-crested (bunodont) hominoid molars. Chimpanzees reportedly spend approximately 5–11 h chewing small (~4 kg) animals[19], and although the carcasses include hide, cartilage and other tough tissues, such lengthy times highlight the challenges of masticating unprocessed meat using low-crested teeth. Consequently, apart from not knowing how much meat early hominins ate, it remains unclear how much adding unprocessed meat to the diet would have affected their ability to chew, especially before cooking became common. Simple cooking methods such as roasting make it easier to chew meat by stiffening muscle fibres and reducing energy dissipation during fracture[20]. Cooking also tends to make plant tissue softer and tenderer, because heat degrades polysaccharides such as pectin and weakens intercellular bonds[20–22]. Another benefit of cooking is to increase the overall energetic yield of both meat and plants[23,24].

It is also important to consider mechanical processing, which is a simpler and older technology. Early *Homo* probably used Lower Palaeolithic tools in at least three ways to process food mechanically. First, rocks can be used to tenderize foods by pounding and grinding, the former of which is observed among chimpanzees[25]. Second, stone flakes are effective for slicing foods into smaller pieces that require fewer chews to consume. Finally, flakes or choppers can be used to remove skin, cartilage, rinds, and other mechanically demanding tissues that are challenging to chew. An added benefit of mechanical processing techniques is to increase net energy yield by breaking down tissues and cell walls, making nutrients more directly accessible to digestion and increasing the surface area to volume ratios of ingested particles[23,24].

Given evidence for meat consumption and the ability to make simple stone tools long before cooking became common, it has long been hypothesized that increased carnivory and the use of Lower Palaeolithic technology made possible selection for smaller teeth and maximum bite forces, as well as other changes in masticatory anatomy evident in *Homo*[6–10]. However, to test these hypotheses it is necessary to compare how mechanical food processing and cooking affect two key masticatory parameters for both meat and plant foods: the muscular effort required for chewing, and how well the food is fragmented (comminuted) before it is swallowed. We therefore measured chewing performance in adult human subjects fed size-standardized samples of meat, as well as USOs, which are hypothesized to have been a particularly important component of the hominin diet[26]. For meat, we used goat, which is relatively tough and therefore more similar to wild game than domesticated beef; for USOs, we used jewel yams, carrots and beets. As described in Methods, these samples were either unprocessed, processed using the two simplest mechanical processing methods available to Lower

[1] Department of Human Evolutionary Biology, Harvard University, 11 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

**Table 1 | Average number of chews and masticatory force used per kcal of USOs and meat**

| Food | | Chews per sample | Applied force per sample (N.s)* | Sample weight (g) | Caloric content (kcal g$^{-1}$) † | Chews per kcal‡ | Applied force per kcal (N.s)‡ |
|---|---|---|---|---|---|---|---|
| USOs§ ($n = 14$) | Unprocessed | 25.2 (±9.2) | 1,105.1 (±539.7) | 2.2 | 0.57 | 20.1 (± 7.4) | 881.3 (±430.4) |
| | Sliced | 26.3 (±10.2) | 1,149.3 (±608.6) | 2.2 | 0.57 | 20.9 (±8.2) | 916.5 (±485.3) |
| | Pounded | 24.2 (±10.0) | 973.2 (±545.5) | 2.1 | 0.57 | 20.2 (±8.4) | 813.0 (±455.8) |
| | Roasted | 22.4 (±8.9) | 870.6 (±489.6) | 2.6 | 0.56 | 15.4 (±6.1) | 597.9 (±336.3) |
| Meat ($n = 10$)‖ | Unprocessed | 40.1 (±19.1) | 1,546.6 (±927.8) | 3.0 | 1.09 | 12.3 (±5.8) | 473.0 (±283.7) |
| | Sliced | 31.2 (±22.0) | 1,099.1 (±1,025.5) | 3.0 | 1.09 | 9.6 (±6.7) | 336.1 (±313.6) |
| | Pounded | 42.1 (±21.7) | 2,033.8 (±1,643.0) | 3.0 | 1.09 | 12.9 (±6.6) | 622.0 (±502.4) |
| | Roasted | 45.3 (±24.8) | 1,924.2 (±850.2) | 3.0 | 1.43 | 10.6 (±5.8) | 448.5 (±198.2) |

Data in brackets are ± 1 s.d.

*Applied masticatory forces were calculated using subject-specific calibration equations that estimate the force–time integral (in N.s) at the M1 from balancing masseter electromyography (EMG) voltage (see Methods).

†Food caloric density was obtained from the US Department of Agriculture (USDA) National Nutrient Database for Standard Reference (http://www.ars.usda.gov/ba/bhnrc/ndl). Sliced and pounded foods were assumed to have the same caloric content as unprocessed foods. Roasted USO data were unavailable, and baked or boiled data were used instead. Unprocessed/sliced/pounded foods: jewel yam = 0.86 kcal g$^{-1}$; red beetroot = 0.43 kcal g$^{-1}$; carrot = 0.41 kcal g$^{-1}$; goat meat = 1.09 kcal g$^{-1}$. Roasted foods: jewel yam = 0.90 kcal g$^{-1}$; red beetroot = 0.44 kcal g$^{-1}$; carrot = 0.35 kcal g$^{-1}$; goat = 1.43 kcal g$^{-1}$.

‡Number of chews and applied masticatory force per kcal of food was calculated by dividing chew number or force per sample by average sample weight and then average caloric density.

§Yam, carrot and beetroot data averaged.

‖Masseter muscle activity was not quantified for one subject, reducing sample size to nine for force per kcal.

Palaeolithic hominins (slicing and pounding), or processed by roasting, the simplest form of cooking.

Comparisons of the number of chews and total applied force required to chew different foods until they were ready to be swallowed (Table 1) indicate that considerably less masticatory effort is required to consume unprocessed meat than USOs. Compared to unprocessed USOs, one kcal of unprocessed meat required on average 39% fewer chews and 46% less force to prepare for swallowing ($P = 0.01$ and $P = 0.02$, respectively). However, the participants we studied were unable to reduce effectively the particle sizes of unprocessed meat through mastication. As Fig. 1 illustrates, even after 40 chews, meat boli were predominately comprised of one large particle (Extended Data Table 1). Therefore, although unprocessed meat requires fewer chews and less force per calorie than USOs, the inability of hominin teeth to break raw, unprocessed meat into small particles probably reduced net energy gain from the food and limited the effectiveness of consuming substantial quantities of unprocessed muscle tissue. This is a conservative estimate since the goat meat samples tested here were already partly processed, lacked cartilage and other mechanically demanding tissue, and were thus relatively unchallenging compared with most of the meat eaten during the Palaeolithic.

Lower Palaeolithic food processing techniques had marked but different effects on the ability to masticate USOs and meat (Table 2, Fig. 1 and Extended Data Tables 1–3). Slicing had no measurable effect on the mastication of USOs, but significantly reduced the average masticatory muscle recruitment used to consume meat by 12.7% per chew ($P < 0.05$) and 31.8% per sample ($P < 0.05$), and also reduced maximum particle size in the comminuted bolus by 40.5% ($P < 0.0001$). In contrast, pounding had no measured effect on the ability to masticate meat, but did reduce the average muscle recruitment used to consume USOs by 4.5% per chew ($P < 0.05$) and 8.7% per sample ($P < 0.05$).

Cooking, whenever it was adopted, would have led to further benefits. Roasted USOs required 14.1% less muscle recruitment per chew ($P < 0.05$) and 22.0% less per sample ($P < 0.05$) compared with unprocessed USOs, but were ready to be swallowed at 82.1% larger particle sizes ($P < 0.01$). Since USOs tend to be tough, force-limited foods[11,20,22], cooking would have substantially reduced hominin peak

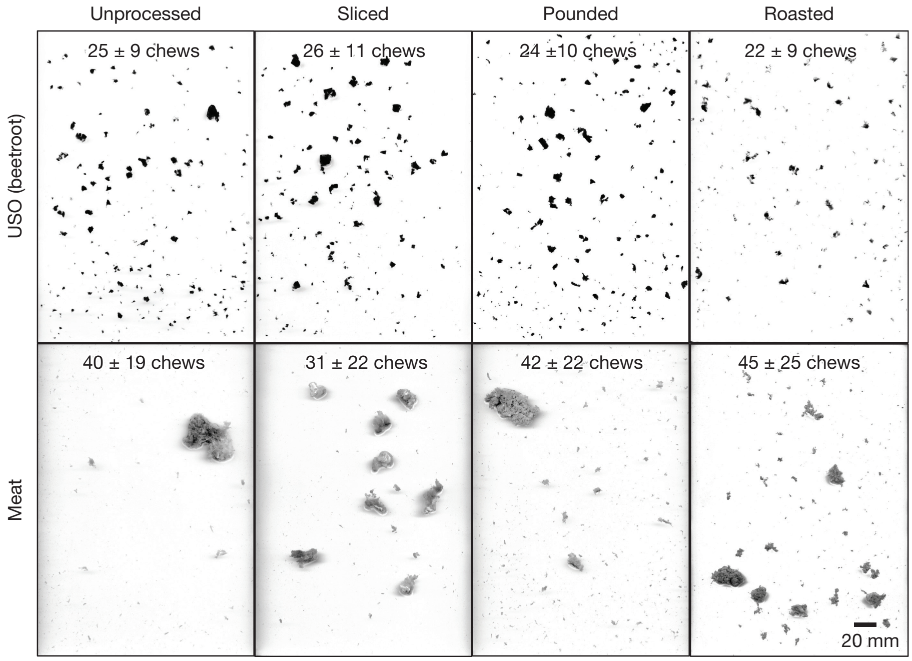


**Figure 1 | Representative samples of chewed meat and USO (beetroot) boli before swallowing.** Particles were dispersed so that they did not overlap. Average number of chews (±1 standard deviation (s.d.)) in parentheses ($n = 10$). Note that when meat is unprocessed (or to a lesser extent pounded) the bolus primarily consists of a single, unfractured food particle. Scale bar, 20 mm.

**Table 2 | Effects of food processing on the mastication of USOs and meat**

| Food | | Muscle recruitment* per chew (% change) | Muscle recruitment* per sample‡ (% change) | Comminution† (% change) |
|---|---|---|---|---|
| USOs | Sliced | NS | NS | NS |
| | Pounded | ↓ 4.5% | ↓ 8.7% | NS |
| | Roasted | ↓ 14.1% | ↓ 22.0% | ↑ 82.1% |
| Meat | Sliced | ↓ 12.7% | ↓ 31.8% | ↓ 40.5% |
| | Pounded | NS | NS | NS |
| | Roasted | ↑ 15.3% | ↑ 32.8% | ↓ 47.1% |

NS, not significant.

*Average percentage change of masticatory muscle recruitment (V.s) resulting from processing USOs (yam, carrot and beetroot data averaged) and meat. $N = 14$ (USOs) and 10 (meat). Only significant changes are shown. Significant changes relative to unprocessed samples are based on 95% confidence intervals greater or less than 0% change, studentized bootstrap (10,000 repeats).

†Participants chewed the food samples (unprocessed and processed beetroots and meat) until they felt the desire to swallow. The size of the largest particle in the chewed bolus was measured and the percentage change resulting from processing calculated. $N = 10$. Only significant changes are shown. Mixed linear models, $P \leq 0.05$.

‡Sum of muscular recruitment per chew used to consume each food sample.

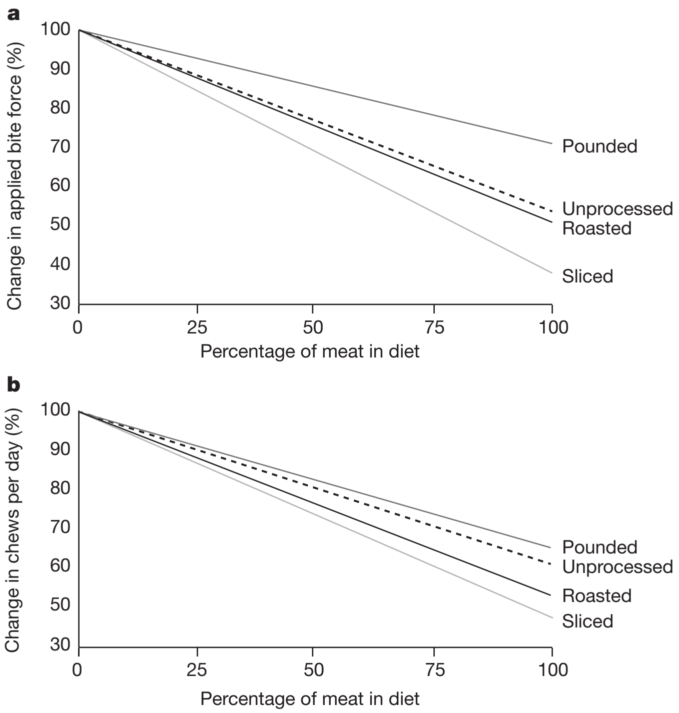


**Figure 2 | Modelled effects of meat and food processing on mastication.** **a**, **b**, Percentage change of applied masticatory force (kN.s) (**a**) and number of chews (**b**) used to consume a 2,000 kcal diet of unprocessed USOs (yam, carrot and beetroot data averaged) and varying amounts of meat that was unprocessed (dashed), sliced (light grey), pounded (dark grey), or roasted (black). Masticatory force was estimated from balancing-side-masseter EMG signals. Applied force and number of chews per kcal were calculated by dividing force or chews per sample by average sample weight and the foods' caloric density. Percentage change of total daily masticatory force and number of chews resulting from the inclusion of unprocessed and processed food was then calculated for diets ranging from 0% to 100% meat.

masticatory effort, in turn reducing selection to maintain large teeth. Assuming that maximum bite force capabilities per chew scale with molar area to the power of 0.9 across primates[3], we can estimate that a 15% reduction in muscle recruitment resulting from roasting USOs would have allowed selection to reduce molar area by approximately 14%; a reduction nearly identical to the approximately 15% smaller post-canines of *H. sapiens* compared to *H. erectus*[4,27].

Roasting also substantially improves the ability to chew meat, although through a different mechanism than USOs. Roasting increased muscular effort by 15.3% per chew ($P < 0.05$) and 32.8% per sample ($P < 0.05$), but decreased the size of the largest particle by 47.1% ($P < 0.0001$), a reduction not significantly different from the effects of slicing meat ($P = 0.81$). In other words, roasted meat required more muscular effort per unit mass to chew, but resulted in a swallowable bolus with smaller particles because of more effective oral fracture.

To model the effects of meat consumption and processing on masticatory forces, we used estimated chewing forces (at the first molar) to predict total daily masticatory force for a hominin consuming 2,000 kcal day$^{-1}$. Although hominins ate many foods, we model the diet as different percentages of USOs and meat (Fig. 2). A diet composed entirely of unprocessed USOs would require approximately 40,000 chews per day. When unprocessed meat is added to the diet, masticatory demands per day decrease by approximately 156 chews and 0.5% of total chewing force for each additional percentage of calories from meat. Thus, if one-third of total calories derive from eating meat (a reasonable estimate based on modern African foraging societies[28]), a hominin would chew approximately 2 million fewer times per year (a 13% reduction) using 15% less total force than on a pure, unprocessed USO diet. If the meat were sliced, then hominins would not only reduce their masticatory effort by more than 2.5 million chews (a 17% reduction) and use 20% less force per year, but they would also swallow meat particles that were approximately 41% smaller, and thus more efficiently digestible[23] (see Tables 1 and 2).

Because the mechanical properties of foods vary depending on many factors such as species and type of portion consumed, further research is necessary to examine additional foods and processing techniques important to human evolution. More research is also needed to quantify the impacts of variations in masticatory morphology on chewing efficiency because dental topography and facial shape affect the relationship between food fracture and chewing effort (for example, sharper cusps increase applied chewing stresses, and relatively shorter jaws increase the mechanical advantage of the adductor muscles). Even so, we speculate that despite the many benefits of cooking for reducing endogenous bacteria and parasites[29], and increasing energy yields[23,24], the reductions in jaw muscle and dental size that evolved by *H. erectus* did not require cooking and would have been made possible by the combined effects of eating meat and mechanically processing both meat and USOs. Specifically, by eating a diet composed of one-third meat, and slicing the meat and pounding the USOs with stone tools before ingestion, early *Homo* would have needed to chew 17% less often and 26% less forcefully. We further surmise that meat eating was largely dependent on mechanical processing made possible by the invention of slicing technology. Meat requires less masticatory force to chew per calorie than the sorts of generally tough plant foods available to early hominins, but the ineffectiveness of hominin molars to break raw meat would have limited the benefits of consuming meat before the invention of stone tools approximately 3.3 Ma. Although recent and contemporary hunter–gatherers are less dependent on stone tools than early *Homo* because they eat mostly cooked meat, many of the oldest tools bear traces of being used to slice meat[9], and the use of tools (now mostly metal knives) to process foods such as meat is well documented ethnographically[30]. This dependency on extra-oral mechanical processing, however, does not apply to other animal-based foods such as marrow, brains and visceral organs that might have been difficult to access without tools, but are easier to chew than muscle. Although it is possible that the masticatory benefits of food processing and carnivory favoured selection for smaller teeth and jaws in *Homo*, we think it is more likely that tool use and meat-eating reduced selection to maintain robust masticatory anatomy, thus permitting selection to decrease facial and dental size for other functions such as speech production, locomotion, thermoregulation, or perhaps even changes in the size and shape of the brain[16]. Whatever selection pressures favoured these shifts, however, they would not have been possible without increased meat consumption combined with food processing technology.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 7 August 2015; accepted 11 January 2016.**
**Published online 9 March 2016.**

1. Aiello, L. C. & Wells, J. C. K. Energetics and the evolution of the genus *Homo*. *Annu. Rev. Anthropol.* **31,** 323–338 (2002).
2. Pontzer, H. Ecological energetics in early *Homo*. *Curr. Anthropol.* **53,** S346–S358 (2012).
3. Eng, C. M., Lieberman, D. E., Zink, K. D. & Peters, M. A. Bite force and occlusal stress production in hominin evolution. *Am. J. Phys. Anthropol.* **151,** 544–557 (2013).
4. McHenry, H. M. Tempo and mode in human evolution. *Proc. Natl Acad. Sci. USA* **91,** 6780–6786 (1994).
5. Aiello, L. C. & Wheeler, P. The expensive-tissue hypothesis: the brain and the digestive-system in human and primate evolution. *Curr. Anthropol.* **36,** 199–221 (1995).
6. Bunn, H. T. in *Evolution of the Human Diet: The Known, the Unknown, and the Unknowable* (ed. Ungar, P.) 191–211 (Oxford Univ. Press, 2007).
7. Domínguez-Rodrigo, M., Pickering, T. R., Semaw, S. & Rogers, M. J. Cutmarked bones from Pliocene archaeological sites at Gona, Afar, Ethiopia: implications for the function of the world's oldest stone tools. *J. Hum. Evol.* **48,** 109–121 (2005).
8. Milton, K. A hypothesis to explain the role of meat-eating in human evolution. *Evol. Anthropol.* **8,** 11–21 (1999).
9. Keeley, L. H. & Toth, N. Microwear polishes on early stone tools from Koobi-Fora, Kenya. *Nature* **293,** 464–465 (1981).
10. Harmand, S. *et al.* 3.3-million-year-old stone tools from Lomekwi 3, West Turkana, Kenya. *Nature* **521,** 310–315 (2015).
11. Lucas, P. *Dental Functional Morphology: How Teeth Work* (Cambridge Univ. Press, 2004).
12. Wrangham, R. W., Jones, J. H., Laden, G., Pilbeam, D. & Conklin-Brittain, N. The raw and the stolen: cooking and the ecology of human origins. *Curr. Anthropol.* **40,** 567–594 (1999).
13. Gowlett, J. & Wrangham, R. W. Earliest fire in Africa: towards the convergence of archaeological evidence and the cooking hypothesis. *Azania Arch. Res. Africa* **48,** 5–30 (2013).
14. Shimelmitz, R. *et al.* 'Fire at will': the emergence of habitual fire use 350,000 years ago. *J. Hum. Evol.* **77,** 196–203 (2014).
15. Larsen, C. S. Animal source foods and human health during evolution. *J. Nutr.* **133** (suppl. 2), 3893S–3897S (2003).
16. Lieberman, D. *The Evolution of the Human Head* (Harvard Press, 2011).
17. Organ, C., Nunn, C. L., Machanda, Z. & Wrangham, R. W. Phylogenetic rate shifts in feeding time during the evolution of *Homo*. *Proc. Natl Acad. Sci. USA* **108,** 14555–14559 (2011).
18. Bramble, D. M. & Lieberman, D. E. Endurance running and the evolution of *Homo*. *Nature* **432,** 345–352 (2004).
19. Wrangham, R. & Conklin-Brittain, N. Cooking as a biological trait. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **136,** 35–46 (2003).
20. Zink, K. D., Lieberman, D. E. & Lucas, P. W. Food material properties and early hominin processing techniques. *J. Hum. Evol.* **77,** 155–166 (2014).
21. Lillford, P. J. Mechanisms of fracture in foods. *J. Texture Stud.* **32,** 397–417 (2001).
22. Dominy, N. J., Vogel, E. R., Yeakel, J. D., Constantino, P. & Lucas, P. W. Mechanical properties of plant underground storage organs and implications for dietary models of early hominins. *Evol. Biol.* **35,** 159–175 (2008).
23. Boback, S. M. *et al.* Cooking and grinding reduces the cost of meat digestion. *Comp. Biochem. Physiol. A Mol. Integr. Physiol.* **148,** 651–656 (2007).
24. Carmody, R. N., Weintraub, G. S. & Wrangham, R. W. Energetic consequences of thermal and nonthermal food processing. *Proc. Natl Acad. Sci. USA* **108,** 19199–19203 (2011).
25. Boesch, C. & Boesch-Achermann, H. *The Chimpanzees of the Tai Forest: Behavioural Ecology and Evolution* (Oxford Univ. Press, 2000).
26. Laden, G. & Wrangham, R. The rise of the hominids as an adaptive shift in fallback foods: plant underground storage organs (USOs) and australopith origins. *J. Hum. Evol.* **49,** 482–498 (2005).
27. Wolpoff, M. H. Posterior tooth size, body size, and diet in South African gracile Australopithecines. *Am. J. Phys. Anthropol.* **39,** 375–393 (1973).
28. Kaplan, H., Hill, K., Lancaster, J. & Hurtado, A. M. A theory of human life history evolution: diet, intelligence, and longevity. *Evol. Anthropol.* **9,** 156–185 (2000).
29. Smith, A. R., Carmody, R. N., Dutton, R. J. & Wrangham, R. W. The significance of cooking for early hominin scavenging. *J. Hum. Evol.* **84,** 62–70 (2015).
30. Gould, R. A. *Living Archaeology* (Cambridge Univ. Press, 1980).

**Acknowledgements** We thank R. Carmody, P. Lucas, J. Shea, T. Smith and R. Wrangham for helpful discussions, and E. Castillo and S. Worthington for statistical guidance. This research was funded by the National Science Foundation (#0925688) and by the American School of Prehistoric Research (Peabody Museum, Harvard University).

**Author Contributions** K.D.Z. and D.E.L. designed the experiments; K.D.Z. collected and analysed the data, with help from D.E.L.; D.E.L. and K.D.Z. co-wrote the paper.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.D.Z. (kzink@oeb.harvard.edu) or D.E.L. (danlieb@fas.harvard.edu).

## METHODS

**Experimental participants.** Three experiments were performed (two on USOs, and one on meat). Experiment number 1 used 14 subjects (7 male, 7 female; aged 29 ± 8 years) to quantify the amount of masticatory muscular effort required to consume USOs. Experiment number 2 used 10 subjects (5 male, 5 female; aged 32 ± 10 years) to quantify comminution (intra-oral food breakdown) of one USO, beetroots. The dark colour of beetroots (but not yams and carrots) provided the colour contrast necessary to image and measure small food particles. Experiment number 3 used 10 male subjects (aged 36 ± 17 years) to quantify both the muscular effort required to consume meat, and how well the subjects were able to comminute the food. All participants had a complete set of permanent teeth with the exception of the third molars (which were variably present), possessed no major cavities, and reported no pain or difficulty during chewing. All experiments were approved by the Harvard Institutional Review Board (IRB), and all subjects provided informed consent before participation. No statistical methods were used to predetermine sample size.

**Food samples.** *USOs.* Organic USOs, jewel yams (*Ipomoea batatas*), carrots (*Daucus carota*), and red beetroots (*Beta vulgaris*), were purchased from a local grocery store. The average fracture toughness of these USOs is approximately 1,060 $\mathrm{J\,m^{-2}}$ (ref. 20), similar to published values from Africa of wild tubers (1,304 $\mathrm{J\,m^{-2}}$), greater than wild bulbs (325 $\mathrm{J\,m^{-2}}$) and corms (265 $\mathrm{J\,m^{-2}}$), but less than wild rhizomes (5,448 $\mathrm{J\,m^{-2}}$) (ref. 22). Each USO was cut into two portions; one was used for the unprocessed samples and the other for the processed samples. Unprocessed, sliced and pounded samples were prepared in a similar manner. First, small bite-sized cubes (13 mm × 13 mm × 13 mm) were cut from the inner medullary region of each USO. Sample dimensions were measured using digital callipers (accuracy ± 0.01 mm). Because of their small size, some of the carrot samples included a small portion of the outer cortex. Sample weight did not differ among the USOs (digital scale, accuracy ± 0.1 g; yam 2.2 ± 0.06 g; carrot 2.3 ± 0.05 g; beetroot 2.2 ± 0.06 g).

After the sample cubes were cut, they were either left unprocessed, or were processed by slicing them into eight smaller 6.5 mm × 6.5 mm × 6.5 mm cubes (sliced samples), or by pounding them six times with a hand-sized rock replica of a Lower Palaeolithic hammerstone (pounded samples). Tenderizing in this manner tended to break the USOs into many relatively large, intact pieces. Roasted samples were created by cutting the USOs into 17-mm-thick slices and then cooking them on a pre-heated tabletop propane grill (Perfect Flame) with the lid open and the gas flow valve set to 'high'. USO slices were roasted for 15 min, and to ensure uniform heating, they were flipped after 7.5 min and rotated every 2.5 min to different positions on the grill surface. Cooking in this manner heated yams to 89.0 ± 2.7 °C, carrots to 78.5 ± 1.1 °C, and beetroots to 78.6 ± 2.2 °C (based on the internal temperatures of 5 slices of each USO; Thermoworks thermometer, accuracy ± 0.1 °C). After cooking, 13 mm × 13 mm × 13 mm cubes were cut from the medullary region of the slices, avoiding the charred surfaces that were in contact with the grill. Cooked cubes were approximately 14% heavier than the unprocessed cubes (cooked yam 2.6 ± 0.05 g; cooked carrot 2.6 ± 0.05 g; cooked beetroot 2.6 ± 0.05 g). All samples were stored in sealed plastic containers at 4 °C and were used within 12 h of processing.

*Meat.* Fresh adult goat carcasses (*Capra aegagrus*; female) were purchased from a local farm (Blood Farms, Groton, Massachusetts) and transported on ice to the Skeletal Biology Laboratory, Harvard University. Neck and epaxial muscles (with little associated fat) were removed using aseptic procedures, sealed in vacuum bags and stored at −20 °C. Although freezing has a slight tenderizing effect[31,32], this step was required by the IRB to perform pathogen tests on the meat before using it in the experiment. The meat was defrosted slowly at 4 °C for approximately 12–24 h before sample preparation. Samples were randomized to include meat from both neck and epaxial muscles. Three-gram samples of meat were cut from defrosted muscles (digital scale, accuracy = 0.1 g). These samples were either left unprocessed, or were cut into eight, approximately equal sized pieces (sliced samples). Pounded samples were created by cutting the muscle into a 50.0 g steak and hitting it 50 times with a replica Lower Palaeolithic hammerstone. Processing in this manner disorganized the muscle fibres, resulting in a 'mashed' appearance, but did not fracture the steak into separate pieces. After tenderizing, 3.0 g samples were cut from the pounded steaks. Roasted samples were created by cooking steaks on the same grill used to cook USOs (see earlier for details). Internal temperature was monitored using a digital thermometer inserted into the steak centre (Thermoworks, accuracy ± 0.1 °C). Steaks were flipped regularly to ensure even heating and were roasted to a final internal temperature equal to medium-well done (slightly pink centre, ~70 °C). On average, cook time was 25.0 ± 5.3 min and water (weight) loss was 26.8 ± 5.6% when roasting in this manner (based on the average of three steaks). After roasting, 3.0 g samples were cut from the steaks, avoiding the charred outer surfaces. All samples were stored in sealed plastic containers at 4 °C and were used within 12 h of processing.

**Order of presentation.** In each of the experiments described later, subjects were presented with triplicate samples of the unprocessed and processed foods. While USO samples were presented in random order, owing to IRB restrictions the cooked meat samples were presented before the unprocessed, sliced and pounded meat samples (the latter three sample types were presented in random order). Additionally, although the subjects were allowed to swallow the USO samples, the risk of foodborne illness precluded swallowing of the non-cooked meat samples. We assessed the potential for bias that non-swallowing might cause by having the subjects chew six samples of cooked meat. Half of the samples were consumed as normal (chewed and swallowed), while the other half were chewed until the subjects felt they would typically swallow, and then spit out. There was no difference in the number of chews used (linear mixed models, $P = 0.65$) or muscle recruitment (linear mixed models, per chew $P = 0.20$, per sample $P = 0.99$). All of the data presented here are based on the cooked meat samples that were not swallowed.

**Muscle recruitment and chewing forces.** For each subject, surface electromyography (EMG) electrodes (Cleartrace, Conmed Corporation) were placed onto the skin overlying the major force-producing muscles of mastication, the right and left temporalis and masseter muscles, and a ground electrode was placed on the back of the non-dominant hand. EMG electrodes were connected to amplifiers (a pre-amplifier and amplifier; MA300 EMG system, Motion Lab Systems) and a PowerLab 16sp A/D board (ADinstruments). All data were collected at 1,000 Hz in LabChart v.7 (ADinstruments). (Temporalis muscle activity was not collected for 3 subjects in the USO experiment, and masseter muscle activity was not collected for 1 subject in the meat experiment.)

**Experimental trials.** After electrode placement, we calibrated each subject's EMGs with force. First, a small, dime-sized Kistler SlimLine force transducer (output voltage calibrated to known forces, $r^2 = 0.99$, for transducer details see later) was placed between the subjects' left first molars. The subjects were then instructed to bite down with sub-maximal force and then release while EMG activity and resulting bite forces were recorded. This procedure was repeated approximately 30 times over a range of bite forces (which were monitored in real time by K.D.Z.). To ensure a comfortable and sterile biting surface, the top and bottom of the transducer was fitted with a thin (2.4 mm) layer of rubber and was loosely covered with a layer of waterproof tape and a sterile plastic sleeve. After wrapping, the transducer was 8.8 mm tall with a diameter of 14.1 mm.

After the calibration trial, subjects were presented with unprocessed and processed foods in randomized order and instructed to chew the samples as normally as possible on only the left side, so that the balancing- and working-side muscles would be readily identifiable. During chewing, the EMG activity of each muscle was recorded. Two sets of analyses were performed: one that assessed the effects of food processing on chewing muscle recruitment, and one that estimated the applied forces necessary to fracture each food. The investigators were not blinded to allocation during experiments and outcome assessment.

**Muscle recruitment analysis.** The EMG signals were processed using custom Matlab codes. Specifically, the data were filtered (Butterworth bandpass; 4th order zero-lag; 60 and 300 Hz frequency cutoffs), rectified, binned with a 5 ms integral reset, and background EMG activity was removed using Thexton's randomization method[33]. Mid-trial swallows, which sometimes occurred during the consumption of the USO samples, were identified by non-uniform patterns of the muscle EMG signals and were omitted from analysis.

For each muscle, the time-integral of the EMG signal was calculated both per chew and per sample. The time-integral EMG data were then normalized within each subject by calculating the relative change in muscular recruitment caused when consuming processed versus unprocessed foods (percentage change = 100 × ((EMG voltage$_{\text{processed food}}$ − EMG voltage$_{\text{unprocessed food}}$)/(EMG voltage$_{\text{unprocessed food}}$)). Sample triplicates were averaged for each subject. Because the data were not normally distributed, we used 95% confidence intervals generated from studentized bootstraps[34] with 10,000 repeats to test whether food processing significantly increased (a positive value) or decreased (a negative value) muscle recruitment. (studentized bootstraps generate confidence intervals based on the resampled distribution of Student's *t*-tests.) EMG data were analysed for each muscle separately, and also with all of the muscles averaged. Similarly, USO data were analysed both for each specific USO (beetroot, carrot and yams), and with all of the USOs pooled together. All calculations were performed in Excel (Microsoft 2007) and R[35].

**Chewing force analysis.** To compare directly the masticatory effort used to chew USOs and meat, we transformed the time-integral EMG data of the balancing-side masseter into estimates of applied chewing force. Although we were not able to estimate the work done by chewing, the time-integral of estimated force is indicative of the total metabolic work done by the muscle, since the percentage of muscle work that generates force is relatively constant (about 25%). Standardization of the EMG signals was necessary because USO and meat samples were different sizes, and EMG signals from different experiments can only be compared when they are normalized. The balancing-side masseter was used because Proeschel and Morneburg[36] found a different EMG–force relationship between isometric bites, such as those used in our calibration experiments, and chewing bites for all major masticatory muscles with the exception of the balancing-side masseter.

To estimate applied chewing forces, subject-specific calibration equations were calculated using the data collected during the calibration trials (see earlier) to transform each subject's muscle recruitment data into chew forces. Specifically, using methods described earlier, we filtered and rectified the balancing-side masseter EMG signal, and calculated the time-integral of the signal for each bite taken in the calibration trial. We then used LabChart v.7 to calculate the time-integral of the force signal used per bite (N.s). Each subject's force data were then regressed against their time-integral EMG data for each bite. Overall, the relationship between the time-integral of the balancing-side masseter EMG and the time-integral of measured bite force was strong and significant: the average $R^2$ ($\pm$ 1 s.d.) for all subject-specific calibration regressions was $0.73 \pm 0.14$; $P \leq 0.001$).

The subject-specific calibration equations generated by the regressions were then used to transform each subject's balancing-side masseter activity per chew into an estimate of applied masticatory force per chew. Total applied masticatory force per sample was then calculated by multiplying the average applied force per chew by the number of chews that a subject used to consume the food.

Finally, the average masticatory force and number of chews used per kcal of each food sample was calculated by dividing by the weight of each sample and the number of calories available per gram of food (see Table 1). All meat samples weighed 3.0 g and USO samples weighed an average of 2.2 g when unprocessed and sliced, 2.1 g when pounded, and 2.6 g when roasted. Food caloric density was obtained from the USDA National Nutrient Database for Standard Reference (http://www.ars.usda.gov/ba/bhnrc/ndl): unprocessed jewel yam = 0.86 kcal $g^{-1}$; unprocessed red beetroot = 0.43 kcal $g^{-1}$; unprocessed carrot = 0.41 kcal $g^{-1}$; unprocessed goat meat = 1.09 kcal $g^{-1}$; baked jewel yam = 0.90 kcal $g^{-1}$; boiled red beetroot = 0.44 kcal $g^{-1}$; boiled carrot = 0.35 kcal $g^{-1}$; roasted goat = 1.43 kcal $g^{-1}$. Caloric data were unavailable for roasted USOs, and baked or boiled USO values were substituted in the calculations. Sliced and pounded foods were assumed to have the same number of calories per gram as their unprocessed counterparts. Yam, carrot and beetroot data were pooled and the average masticatory force per kcal of USO was calculated. A two-tailed Mann–Whitney *U*-test was used to assess whether the number of chews and masticatory force used to eat a kilocalorie of food differed between USOs and meat. All calculations were performed in Excel (Microsoft 2007) and StatView (SAS Institute). Significance was set to $P \leq 0.05$.

It should be noted that the caloric values used in these calculations are based on the Atwater system, which calculates food energy as the total available energy minus the indigestible components. This system assumes a standard digestibility, however, and also fails to take into account other key variables, such as the cost of digestion, which is lower in processed foods[23,24]. Therefore, these caloric data probably under-report the net energy gained from processed foods.

**Comminution.** Subjects were instructed to chew the meat and USO (beetroot) samples on the left side of their mouth until they felt that they would typically swallow. At this point they stopped chewing and the food bolus was collected in 50 ml tubes and stored in ~50% ethanol for no more than 8 days before image analysis.

**Particle size analysis.** Comminuted boli were dispersed onto a transparent plastic tray fitted onto an Epson perfection v500 flatbed scanner. Food particles comprising each bolus were easily separated using water, and were arranged so that the particles did not touch one another and to maximize surface area contact with the tray. Particles were then scanned to create a 400 dpi grey-scale image against a white background. Images were viewed and measured in iVision v.4 (BioVision Technologies).

Comminution effectiveness was quantified as the two-dimensional surface area of the largest particle of food within the chewed bolus. We use this variable rather than average particle size because the chewed boli of unprocessed meat were predominantly composed of a single large particle, making average size uninformative (see Fig. 1). In most instances, the largest particle in a chewed meat bolus was readily identifiable in the scanned images. Using the drawing tool, the pixels comprising the largest particle were manually transformed into the measurement colour (green), and the total two-dimensional surface area ($mm^2$) of the particle was then quantified based on the number of coloured pixels. In some samples, multiple particles had to be measured to locate the largest particle. In contrast to meat, the comminuted USO samples contained a large number of similarly sized particles, and it was not possible to discern the largest particle simply by viewing the scanned images. Therefore, all of the particles that made up the sample were measured. To do this, the scanned image was thresholded so that every coloured pixel with a value ranging from 0 to 230 (pure black to very light grey, respectively) was transformed into the measurement colour (green). (Preliminary tests indicated that thresholding to 230 was the boundary between very small, light particles and shadows resulting from the scanner's moving light source.) After thresholding, the image was reviewed and digitally cleaned by hand if needed. The surface area of every individual food particle was measured by quantifying the number of green pixels comprising the particle (a single particle was defined as the sum of all green pixels in contact). For consistency, we report only data on the size of the largest particle in the chewed USO boli, which correlated strongly with average particle size ($r = 0.73$; $P < 0.0001$) (see Extended Data Table 1).

Triplicates of each sample type were averaged, and the size of the largest particle in raw and processed comminuted samples was compared using linear mixed models, a type of model that estimates separate intercepts for each subject[37]. All calculations were performed in Excel (Microsoft 2007) and R[35]. Significance was set to $P \leq 0.05$. Measurement precision was quantified by measuring the bolus of one randomly chosen sample (unprocessed meat) five times. The standard deviation of the resulting measurements (1.4 $mm^2$) was 0.2% that of the average particle area (542.6 $mm^2$). The maximum difference between any two repeats was 0.5% of the average.

31. Lagerstedt, A., Enfält, L., Johansson, L. & Lundström, K. Effect of freezing on sensory quality, shear force and water loss in beef *M. longissimus dorsi*. *Meat Sci.* **80,** 457–461 (2008).
32. Vieira, C., Diaz, M. T., Martínez, B. & García-Cachán, M. D. Effect of frozen storage conditions (temperature and length of storage) on microbiological and sensory quality of rustic crossbred beef at different states of ageing. *Meat Sci.* **83,** 398–404 (2009).
33. Thexton, A. J. A randomisation method for discriminating between signal and noise recordings of rhythmic electromyographic activity. *J. Neurosci. Methods* **66,** 93–98 (1996).
34. Carpenter, J. & Bithell, J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.* **19,** 1141–1164 (2000).
35. R Development Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2014).
36. Proeschel, P. A. & Morneburg, T. Task-dependence of activity/bite-force relations and its impact on estimation of chewing force from EMG. *J. Dent. Res.* **81,** 464–468 (2002).
37. Bolker, B. M. *et al.* Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* **24,** 127–135 (2009).

**Extended Data Table 1 | The number and size of food particles contained within chewed USO (beetroot) and meat boli at 'swallow'**

| Food | | Particle Size ($mm^2$) | | | Total Surface Area ($mm^2$) | Number of Particles |
|---|---|---|---|---|---|---|
| | | *Largest* | *Median* | *Average* | | |
| USOs (Beets) | *Unprocessed* | 38.52 (26.39) | 0.22 (0.09) | 1.23 (0.41) | 1593.28 (423.47) | 1434.10 (635.19) |
| | *Sliced* | 33.05 (17.49) | 0.26 (0.10) | 1.37 (0.65) | 1594.36 (399.07) | 1367.60 (637.73) |
| | *Pounded* | 31.87 (17.04) | **0.29 (0.07)** | **1.49 (0.58)** | 1567.62 (412.09) | 1159.50 (454.40) |
| | *Roasted* | **57.36 (24.90)** | **0.30 (0.08)** | **2.08 (0.62)** | 1922.87 (406.84) | 1061.30 (653.39) |
| Meat | *Unprocessed* | 697.3 (207.9) | | | | |
| | *Sliced* | **363.6 (108.8)** | | | | |
| | *Pounded* | 759.5 (136.7) | | | | |
| | *Roasted* | **378.8 (183.3)** | | | | |

Subjects chewed unprocessed and processed USOs (beetroots) and meat until they felt that they would typically swallow. The chewed bolus was collected and the two-dimensional surface area of the particles comprising the bolus measured. $N = 10$. Average values are shown. One standard deviation in parentheses. Bold values indicate a significant difference between processed and unprocessed foods. Mixed linear models, $P \leq 0.05$.

**Extended Data Table 2 | Average percentage change of chewing muscle recruitment per chew when masticating size-standardized processed USOs and meat, relative to unprocessed samples**

| Food | | **Balancing Side** (% Change) | | **Working Side** (% Change) | | **Muscle Average**<br>(% Change) |
|---|---|---|---|---|---|---|
| | | *Masseter* | *Temporalis* | *Masseter* | *Temporalis* | |
| Yam<br>*(N=14[†])* | *Sliced* | 0.1 (9.5)<br>*95% CI: 6 to -5* | 3.6 (8.9)<br>*95% CI: 9 to -4* | -0.3 (10.7)<br>*95% CI: 5 to -8* | 3.4 (9.4)<br>*95% CI: 12 to -1* | **0.8 (5.6)**<br>*95% CI: 3 to -4* |
| | *Pounded* | -4.9 (10.1)<br>*95% CI: 2 to -10* | -2.2 (12.0)<br>*95% CI: 10 to -8* | -0.9 (19.8)<br>*95% CI: 14 to -10* | 0.2 (10.7)<br>*95% CI: 14 to -5* | **-2.3 (9.9)**<br>*95% CI: 7 to -7* |
| | *Roasted* | -24.2 (14.8)<br>*95% CI: -14 to -32* | -32.1 (13.6)<br>*95% CI: -22 to -44* | -25.7 (13.0)<br>*95% CI: -15 to -32* | -30.7 (15.4)<br>*95% CI: -15 to -39* | **-27.3 (14.7)**<br>*95% CI: -15 to -34* |
| Carrot<br>*(N=14[†])* | *Sliced* | -0.2 (8.9)<br>*95% CI: 6 to -5* | -1.4 (10.3)<br>*95% CI: 5 to -9* | -0.6 (10.0)<br>*95% CI: 5 to -7* | 4.0 (14.3)<br>*95% CI: 16 to -4* | **0.7 (8.8)**<br>*95% CI: 6 to -4* |
| | *Pounded* | -8.7 (13.3)<br>*95% CI: -1 to -17* | -6.1 (10.3)<br>*95% CI: 0.2 to -15* | -3.5 (8.8)<br>*95% CI: 1 to -11* | -2.3 (13.1)<br>*95% CI: 7 to -11* | **-4.4 (8.7)**<br>*95% CI: -0.1 to -11* |
| | *Roasted* | -7.3 (7.9)<br>*95% CI: -1 to -11* | -9.7 (8.2)<br>*95% CI: -5 to -20* | -13.3 (9.9)<br>*95% CI: -8 to -20* | -6.2 (9.3)<br>*95% CI: -2 to -17* | **-9.7 (6.2)**<br>*95% CI: -6 to -14* |
| Beet<br>*(N=14[†])* | *Sliced* | -3.1 (14.9)<br>*95% CI: 5 to -13* | 2.4 (7.8)<br>*95% CI: 7 to -3* | 8.3 (28.8)<br>*95% CI: 53 to -3* | -2.8 (8.7)<br>*95% CI: 4 to -8* | **-0.2 (8.3)**<br>*95% CI: 5 to -4* |
| | *Pounded* | -11.9 (15.5)<br>*95% CI: -3 to -21* | -7.0 (10.1)<br>*95% CI: -1 to -15* | -4.4 (8.9)<br>*95% CI: 1 to -9* | -7.6 (7.2)<br>*95% CI: -3 to -14* | **-6.8 (7.6)**<br>*95% CI: -2 to -11* |
| | *Roasted* | -4.1 (8.0)<br>*95% CI: 0.2 to -9* | -2.3 (6.1)<br>*95% CI: 1 to -7* | -6.3 (14.8)<br>*95% CI: 0.4 to -20* | -4.8 (7.5)<br>*95% CI: -0.5 to -11* | **-5.4 (6.8)**<br>*95% CI: -2 to -10* |
| USO<br>Average | *Sliced* | -1.1 (7.6)<br>*95% CI: 3 to -8* | 1.5 (4.4)<br>*95% CI: 4 to -3* | 2.4 (10.2)<br>*95% CI: 12 to -2* | 1.5 (5.4)<br>*95% CI: 5 to -2* | **0.5 (3.4)**<br>*95% CI: 2 to -2* |
| | *Pounded* | -8.5 (9.1)<br>*95% CI: -3 to -13* | -5.1 (8.0)<br>*95% CI: 1 to -10* | -3.0 (7.9)<br>*95% CI: 3 to -7* | -3.2 (7.4)<br>*95% CI: 2 to -8* | **-4.5 (6.1)**<br>*95% CI: -0.01 to -7* |
| | *Roasted* | -11.9 (6.9)<br>*95% CI: -7 to -15* | -14.7 (6.9)<br>*95% CI: -10 to -19* | -15.1 (10.0)<br>*95% CI: -10 to -23* | -13.9 (7.6)<br>*95% CI: -8 to -19* | **-14.1 (6.8)**<br>*95% CI: -9 to -18* |
| Meat<br>*(N=10[‡])* | *Sliced* | -13.8 (15.0)<br>*95% CI: 3 to -24* | -14.3 (14.3)<br>*95% CI: -5 to -26* | -12.0 (9.9)<br>*95% CI: -6 to -22* | -11.1 (10.3)<br>*95% CI: -3 to -18* | **-12.7 (10.1)**<br>*95% CI: -5 to -20* |
| | *Pounded* | 13.5 (12.9)<br>*95% CI: 23 to 2* | 8.6 (16.5)<br>*95% CI: 23 to -2* | 8.3 (14.5)<br>*95% CI: 18 to -6* | 5.2 (11.5)<br>*95% CI: 14 to -3* | **7.9 (12.6)**<br>*95% CI: 18 to -1* |
| | *Roasted* | 24.4 (21.3)<br>*95% CI: 42 to 9* | 12.0 (18.2)<br>*95% CI: 32 to 1* | 17.7 (24.3)<br>*95% CI: 51 to 3* | 12.5 (17.2)<br>*95% CI: 38 to 4* | **15.3 (18.1)**<br>*95% CI: 40 to 6* |

For each food, muscle recruitment percentage change = 100 × ((EMG voltage per chew$_{\text{processed food}}$ − EMG voltage per chew$_{\text{unprocessed food}}$)/(EMG voltage per chew$_{\text{unprocessed food}}$)). One standard deviation in parentheses. Significant changes relative to unprocessed samples are shaded grey based on 95% confidence intervals (CI) greater or less than 0% change, studentized bootstrap (10,000 repeats).

[*] The temporalis muscle was not collected from 3 subjects, reducing samples size to 11 for this muscle.

[†] The masseter muscle was not collected from 1 subject, reducing sample size to 9 for this muscle.

**Extended Data Table 3 | Average percentage change of chewing muscle recruitment per sample when masticating size-standardized processed USOs and meat, relative to unprocessed samples**

| Food | | **Balancing Side** (% Change) | | **Working Side** (% Change) | | **Muscle Average**<br>(% Change) |
|---|---|---|---|---|---|---|
| | | *Masseter* | *Temporalis* | *Masseter* | *Temporalis* | |
| Yam<br>*(N=14[‡])* | *Sliced* | 3.1 (10.5)<br>*95% CI: 8 to -5* | 6.6 (11.8)<br>*95% CI: 18 to 1* | 2.4 (10.0)<br>*95% CI: 10 to -2* | 6.7 (13.7)<br>*95% CI: 17 to -2* | **3.8 (9.4)**<br>*95% CI: 9 to -2* |
| | *Pounded* | -13.1 (15.8)<br>*95% CI: -2 to -21* | -11.1 (17.0)<br>*95% CI: 4 to -21* | -10.0 (18.3)<br>*95% CI: -1 to -23* | -9.0 (17.3)<br>*95% CI: 3 to -19* | **-10.9 (16.1)**<br>*95% CI: -1 to -19* |
| | *Roasted* | -42.0 (20.2)<br>*95% CI: -27 to -52* | -49.9 (16.8)<br>*95% CI: -36 to -60* | -43.6 (19.1)<br>*95% CI: -30 to -53* | -48.3 (19.8)<br>*95% CI: -25 to -59* | **-44.3 (20.1)**<br>*95% CI: -28 to -54* |
| Carrot<br>*(N=14[‡])* | *Sliced* | 2.4 (12.1)<br>*95% CI: 10 to -4* | -0.8 (10.5)<br>*95% CI: 6 to -8* | 2.7 (15.9)<br>*95% CI: 16 to -4* | 4.7 (13.5)<br>*95% CI: 16 to -3* | **3.5 (13.2)**<br>*95% CI: 15 to -2* |
| | *Pounded* | -9.9 (16.4)<br>*95% CI: 2 to -18* | -10.5 (12.4)<br>*95% CI: -1 to -17* | -4.1 (17.8)<br>*95% CI: 8 to -13* | -6.6 (14.3)<br>*95% CI: 3 to -16* | **-5.4 (14.9)**<br>*95% CI: 5 to -13* |
| | *Roasted* | -13.6 (10.1)<br>*95% CI: -7 to -19* | -17.3 (14.5)<br>*95% CI: -7 to -26* | -17.6 (16.0)<br>*95% CI: -5 to -25* | -13.5 (16.9)<br>*95% CI: 1 to -23* | **-15.0 (13.0)**<br>*95% CI: -5 to -21* |
| Beet<br>*(N=14[‡])* | *Sliced* | 4.1 (27.3)<br>*95% CI: 25 to -8* | 11.6 (23.9)<br>*95% CI: 33 to -0.1* | 14.6 (35.4)<br>*95% CI: 83 to -1* | 5.8 (23.7)<br>*95% CI: 30 to -6* | **6.8 (22.8)**<br>*95% CI: 27 to -2* |
| | *Pounded* | -14.8 (17.2)<br>*95% CI: -4 to -25* | -9.5 (16.1)<br>*95% CI: 1 to -20* | -8.1 (13.1)<br>*95% CI: 3 to -14* | -9.8 (14.8)<br>*95% CI: -0.4 to -20* | **-9.9 (13.3)**<br>*95% CI: -1 to -17* |
| | *Roasted* | -5.8 (14.0)<br>*95% CI: 2 to -14* | -4.0 (13.2)<br>*95% CI: 6 to -12* | -7.1 (22.2)<br>*95% CI: 4 to -22* | -6.6 (14.8)<br>*95% CI: 5 to -15* | **-6.7 (14.9)**<br>*95% CI: 2 to -15* |
| *USO Average* | *Sliced* | 3.2 (12.8)<br>*95% CI: 10 to -4* | 5.8 (11.7)<br>*95% CI: 15 to -1* | 6.6 (14.2)<br>*95% CI: 21 to 0.3* | 5.7 (13.3)<br>*95% CI: 17 to -2* | **4.7 (11.0)**<br>*95% CI: 11 to -1* |
| | *Pounded* | -12.6 (12.5)<br>*95% CI: -4 to -19* | -10.4 (12.2)<br>*95% CI: 3 to -17* | -7.4 (10.0)<br>*95% CI: 0.3 to -12* | -8.5 (11.5)<br>*95% CI: -0.1 to -16* | **-8.7 (10.8)**<br>*95% CI: -0.4 to -14* |
| | *Roasted* | -20.5 (9.3)<br>*95% CI: -15 to -26* | -23.8 (10.0)<br>*95% CI: -17 to -31* | -22.8 (13.5)<br>*95% CI: -16 to -32* | -22.8 (11.4)<br>*95% CI: -15 to -30* | **-22.0 (10.5)**<br>*95% CI: -16 to -28* |
| Meat<br>*(N=10[§])* | *Sliced* | -29.2 (36.1)<br>*95% CI: 5 to -52* | -33.3 (30.4)<br>*95% CI: -8 to -52* | -29.3 (31.3)<br>*95% CI: -2 to -51* | -30.7 (31.9)<br>*95% CI: -2 to -49* | **-31.8 (31.2)**<br>*95% CI:-5 to -50* |
| | *Pounded* | 28.1 (42.2)<br>*95% CI: 58 to -7* | 19.8 (44.1)<br>*95% CI: 53 to -12* | 24.2 (40.2)<br>*95% CI: 52 to -15* | 15.4 (38.6)<br>*95% CI: 45 to -12* | **18.7 (40.9)**<br>*95% CI: 48 to -12* |
| | *Roasted* | 42.8 (56.9)<br>*95% CI: 100 to 6* | 28.4 (52.0)<br>*95% CI: 85 to 1* | 37.4 (61.2)<br>*95% CI: 98 to -2* | 29.5 (47.8)<br>*95% CI: 75 to 1* | **32.8 (51.7)**<br>*95% CI: 82 to 3* |

Chewing muscle recruitments per sample were calculated as the sum of muscular recruitment per chew used to consume each food sample. For each food, muscle recruitment percentage change = 100 × ((EMG voltage per sample$_{\text{processed food}}$– EMG voltage per sample $_{\text{unprocessed food}}$)/(EMG voltage per sample $_{\text{unprocessed food}}$)). One standard deviation in parentheses. Significant changes relative to unprocessed samples are shaded grey based on 95% confidence intervals (CI) greater or less than 0% change, studentized bootstrap (10,000 repeats).

[*]The temporalis muscle was not collected from 3 subjects, reducing samples size to 11 for this muscle.

[†]The masseter muscle was not collected from 1 subject, reducing sample size to 9 for this muscle.

doi:10.1038/nature17405

# Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins

Matthias Meyer[1], Juan-Luis Arsuaga[2,3], Cesare de Filippo[1], Sarah Nagel[1], Ayinuer Aximu-Petri[1], Birgit Nickel[1], Ignacio Martínez[2,4], Ana Gracia[2,4], José María Bermúdez de Castro[5,6], Eudald Carbonell[7,8], Bence Viola[9], Janet Kelso[1], Kay Prüfer[1] & Svante Pääbo[1]

**A unique assemblage of 28 hominin individuals, found in Sima de los Huesos in the Sierra de Atapuerca in Spain, has recently been dated to approximately 430,000 years ago[1]. An interesting question is how these Middle Pleistocene hominins were related to those who lived in the Late Pleistocene epoch, in particular to Neanderthals in western Eurasia and to Denisovans, a sister group of Neanderthals so far known only from southern Siberia. While the Sima de los Huesos hominins share some derived morphological features with Neanderthals, the mitochondrial genome retrieved from one individual from Sima de los Huesos is more closely related to the mitochondrial DNA of Denisovans than to that of Neanderthals[2]. However, since the mitochondrial DNA does not reveal the full picture of relationships among populations, we have investigated DNA preservation in several individuals found at Sima de los Huesos. Here we recover nuclear DNA sequences from two specimens, which show that the Sima de los Huesos hominins were related to Neanderthals rather than to Denisovans, indicating that the population divergence between Neanderthals and Denisovans predates 430,000 years ago. A mitochondrial DNA recovered from one of the specimens shares the previously described relationship to Denisovan mitochondrial DNAs, suggesting, among other possibilities, that the mitochondrial DNA gene pool of Neanderthals turned over later in their history.**

When modern humans spread out of Africa and the Near East some 75,000–50,000 years ago, at least two archaic hominin groups, Neanderthals and Denisovans, inhabited Eurasia. While Neanderthals are known from an abundant fossil record in Europe and western and central Asia, Denisovan remains are currently only known from the Altai Mountains in southern Siberia[3,4]. However, Denisovan ancestry is detected in present-day human populations from Oceania, mainland Asia and in Native Americans[5], suggesting that they were once more widespread. High-quality genome sequences recovered from one Neanderthal and one Denisovan show that they were more closely related to each other than to modern humans[6,7] and that they diverged from a common ancestral population between 381,000 and 473,000 years ago[7] if a mutation rate of $0.5 \times 10^{-9}$ per site per year is used.

The Middle Pleistocene fossils from Sima de los Huesos (SH) are relevant for the question of when and where the ancestral populations of Neanderthals and Denisovans lived, but their relationship to these later archaic groups is unclear. They share some derived dental and cranial features with Late Pleistocene Neanderthals, for example, a midfacial prognathism and some aspects of the supraorbital torus, the occipital bone and the glenoid cavity[1,8]. In apparent contrast to this, the mitochondrial (mt)DNA determined from one SH individual is more similar to an mtDNA ancestral to Denisovan than to Neanderthal mtDNAs[2]. However, the mtDNA is inherited as a single unit from mothers to offspring and does not necessarily reflect the overall relationship of individuals and populations. To clarify the relationships of the SH hominins to Neanderthals and Denisovans, we therefore set out to retrieve nuclear DNA from SH hominins. However, DNA preservation in these fossils is poor owing to their great age. Femur XIII, from which the SH mtDNA genome was sequenced, contains only small amounts of highly degraded endogenous DNA (30–45 base pairs (bp)) in a large excess of microbial DNA. To reconstruct its mtDNA genome, almost 2 g of bone had to be used to produce DNA libraries from which mtDNA fragments were isolated by hybridization capture. Furthermore, because of the presence of modern human DNA contamination, putatively endogenous sequences had to be identified on the basis of the presence of C to T substitutions that accumulate at the ends of DNA fragments over time owing to cytosine deamination[9], which are largely absent in recent human DNA that contaminates fossils[10,11].

To retrieve nuclear DNA sequences from femur XIII, we generated approximately 2.6 billion sequence reads from the library with the highest frequency of terminal C to T substitutions (library A2021 (ref. 2)). In addition, between 600 million and 900 million reads were collected from each of four new specimens that were recovered from the site for molecular analyses (Extended Data Table 1). These were an incisor (AT-5482), a femur fragment (AT-5431), a molar (AT-5444) and a scapula (AT-6672).

In addition to sequencing random fragments from these specimens, we also isolated mtDNA fragments from the four new specimens by hybridization capture. Between 1,419 and 3,742 unique mtDNA fragments of 30 bp or longer were retrieved (Extended Data Table 2). To investigate whether they represented endogenous DNA or present-day human contamination, we determined the frequency of C to T substitutions relative to the human mitochondrial genome at each position in the fragments. The fragments from femur AT-5431 carry 44% C to T substitutions at the 5′ ends and 41% at the 3′ ends, compatible with the presence of endogenous ancient mtDNA. The other three specimens do not show discernible evidence of deamination-induced substitutions. Because there were too few DNA fragments to reconstruct the complete mtDNA genome of femur AT-5431, we restricted further analyses to 'diagnostic' positions in the mtDNA genome where each lineage in the mtDNA tree differs from the other hominin lineages and from the chimpanzee. At positions where modern humans differ from Neanderthals, Denisovans, SH femur XIII and the chimpanzee, 41% (17 out of 41) of the mtDNA fragments share the modern human state, indicating that they are derived from present-day human contamination

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. [2]Centro de Investigación Sobre la Evolución y Comportamiento Humanos, Universidad Complutense de Madrid–Instituto de Salud Carlos III, 28029 Madrid, Spain. [3]Departamento de Paleontología, Facultad de Ciencias Geológicas, Universidad Complutense de Madrid, 28040 Madrid, Spain. [4]Área de Paleontología, Departamento de Geografía y Geología, Universidad de Alcalá, Alcalá de Henares, 28871 Madrid, Spain. [5]Centro Nacional de Investigación sobre la Evolución Humana, Paseo Sierra de Atapuerca, 09002 Burgos, Spain. [6]Department of Anthropology, University College London, 14 Taviton Street, London WC1H 0BW, UK. [7]Institut Català de Paleoecologia Humana i Evolució Social, C/Marcel·lí Domingo s/n (Edifici W3), Campus Sescelades, 43007 Tarragona, Spain. [8]Àrea de Prehistòria, Departament d'Història i Història de l'Art, Universitat Rovira i Virgili, Facultat de Lletres, Avinguda de Catalunya, 35, 43002 Tarragona, Spain. [9]Department of Anthropology, University of Toronto, 19 Russell Street, Toronto, Ontario M5S 2S2, Canada.

**Table 1 | Characteristics of the nuclear sequence alignments obtained from the five SH specimens**

| Specimen | Sequences ≥35 bp mapped (%) | C to T substitution frequency (%) | | | | Aligned sequence in putatively deaminated fragments (bp) | Present-day human contamination (%)* | |
|---|---|---|---|---|---|---|---|---|
| | | | | 'Conditional' | | | | |
| | | 5′ end | 3′ end | 5′ end | 3′ end | | All fragments | Putatively deaminated fragments |
| Femur XIII (AT2944) | 0.02 | 5.3 | 8.2 | 32.6 | 41.9 | 225,329 | 88 | 63 |
| Incisor (AT-5482) | 0.25 | 12.9 | 15.9 | 60.4 | 58.5 | 2,015,167 | 86 | 21 |
| Femur frag. (AT-5431) | 0.09 | 21.5 | 22.3 | 63.6 | 55.8 | 1,186,442 | 76 | 18 |
| Molar (AT-5444) | 0.02 | 10.6 | 11.8 | 63.1 | 53.4 | 188,974 | 88 | 0 |
| Scapula (AT-6672) | 0.03 | 0.6 | 1.0 | 4.4 | 10.0 | 37,524 | 98 | 100 |

*Present-day human contamination is estimated here as the percentage of SH sequences that share alleles derived in 90% or more of present-day humans and ancestral in the two archaic high-coverage genomes and other primates.

(Extended Data Fig. 1). In contrast, among the five fragments that show evidence for deamination, none shares the human-derived state. Among the eight putatively deaminated fragments that overlap positions with variants specific to Denisovans and to femur XIII, all eight carry the derived variants present in both lineages. In addition, three out of five fragments carry variants specific to femur XIII only. In contrast, of nine mtDNA fragments overlapping positions diagnostic for Neanderthals, none carry the Neanderthal variants. We thus conclude that the mtDNA of femur AT-5431 is most closely related to the mtDNA of femur XIII.

The vast majority of endogenous DNA in the SH fossils is degraded to a size below 45 bp. To maximize the yield of DNA fragments we have therefore used fragments as short as 30 bp when reconstructing mtDNAs from these specimens[2,12], whereas DNA analyses from other archaic hominins have been restricted to fragments of 35 bp or longer[6,7]. We explored whether it might be possible to use DNA fragments as short as 30 bp to study also the nuclear genome in SH specimens; however, this resulted in 9–67% of the aligned DNA fragments appearing to be of microbial rather than of hominin origin. In contrast, no spurious alignments were detected at a length cut-off of ≥35 bp (Supplementary Information, section 1). Using the latter cut-off, between 0.02% and 0.25% of the DNA sequences determined from the fossils align to the human reference genome (Table 1). With the exception of the scapula, all specimens show C to T substitution frequencies between 5% and 22% at the terminal alignment positions. When conditioned on C to T substitutions at the other ends of fragments, they increase to between 53% and 64% for the incisor, the femur fragment and the molar, to 33% and 42% for femur XIII, and to 4% and 10% for the scapula (Table 1 and Extended Data Fig. 2), indicating that all five specimens carry mixtures of highly deaminated endogenous nuclear DNA and less deaminated human contamination (Supplementary Information, section 2).

Owing to the extremely small amounts of data available, assessment of nuclear DNA contamination cannot be achieved using existing approaches that require multi-fold coverage in at least parts of the nuclear genome[13,14]. We therefore used two alternative approaches to obtain estimates of contamination for the nuclear sequences (Supplementary Information, sections 2 and 3). The first approach, which compares deamination signals in all sequences to those carrying a C to T substitution at the opposing end, estimates human contamination to be >63% in all five specimens. The second approach estimates contamination as the percentage of sequences sharing the modern human state at sites where 90% or more of present-day humans differ from the chimpanzee and the two high-coverage archaic genomes. The contamination estimates from this approach are similarly high (Table 1), but decrease to 21% or less in three of the specimens (femur AT-5431, the incisor and the molar) when filtering for sequences showing terminal C to T substitutions indicative of deamination. Disregarding fragments without evidence for deamination, the amount of nuclear DNA sequence retrieved varies between 189 kb and 2.0 Mb for these three specimens (Table 1), all of which are male as inferred from the sequence coverage of chromosome X and the autosomes (Extended Data Fig. 3). The sex of the scapula and femur XIII cannot be confidently determined as a result of the high levels of present-day human contamination and the limited amount of data available.

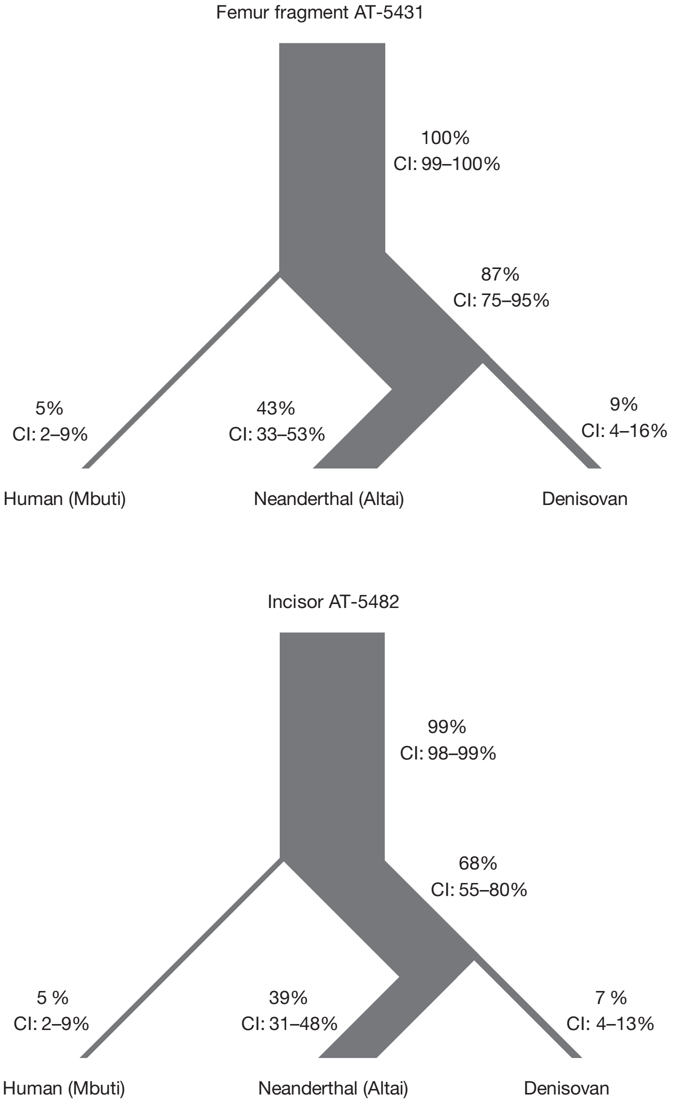


**Figure 1 | Percentage of derived alleles shared between the SH specimen and the human, Neanderthal and Denisovan genomes.** Ninety-five per cent binomial confidence intervals (CI) are indicated. The thickness of the branches is scaled by the extent of derived allele sharing. See Extended Data Fig. 4 for the total number of informative positions identified in the nuclear genome and Extended Data Table 3 for the number of sequences overlapping these positions.

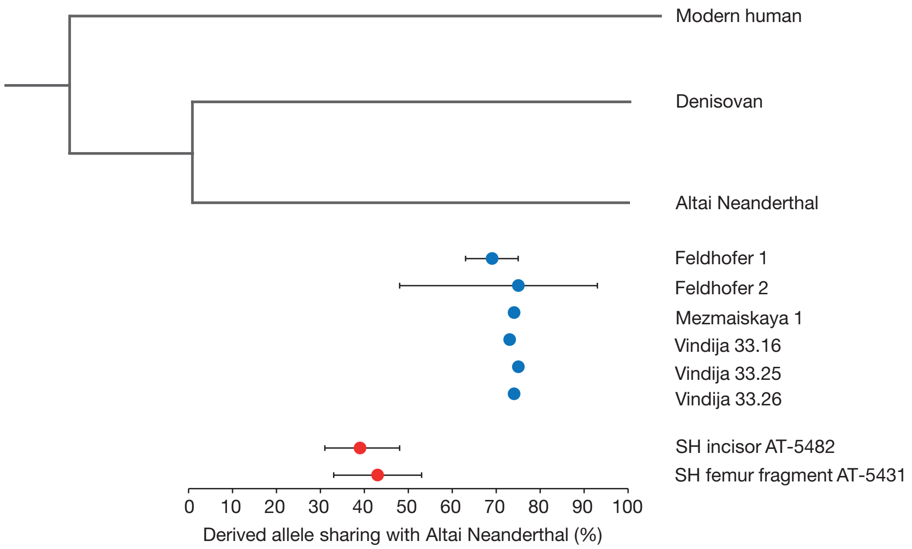


**Figure 2 | Sharing of derived alleles with the Altai Neanderthal.** Error bars, 95% confidence intervals.

To investigate how the SH hominins are related to modern humans, Neanderthals and Denisovans, we used the high-quality genome sequences of the Altai Neanderthal, the Denisovan finger bone and a present-day human individual from Africa (Mbuti, HGDP00982) to identify positions where one or more of these three genomes differ from those of the chimpanzee and other primates (bonobo, gorilla, orangutan, rhesus macaque) (Extended Data Fig. 4). We then estimated the percentages of all informative positions covered in each specimen that share the derived state for each branch in the tree relating the three genome sequences. For the femur fragment AT-5431 and the incisor, we find that 87% and 68%, respectively, of the positions on the common Neanderthal and Denisovan branch carry derived alleles; that 43% and 39%, respectively, of positions on the Neanderthal branch carry derived alleles; while 9% and 7%, respectively, on the Denisovan branch do so (Fig. 1). This indicates that the SH hominins are related to the ancestors of Neanderthals rather than Denisovans. The fraction of derived alleles shared with the Neanderthal genome is between two- and three-fold smaller when sequences without terminal C to T substitutions are also included (Extended Data Table 3), confirming that the signal linking the SH hominins to Neanderthals is derived from endogenous DNA fragments. These results are stable when present-day human individuals other than the Mbuti are used in the analysis (Extended Data Table 4). Similar to femur AT-5431 and the incisor, the molar also shows a greater sharing of derived alleles with the Neanderthal than the Denisovan genome, although not statistically significantly so, probably because of the small amount of data available (Extended Data Fig. 5). By comparison, the fraction of derived alleles shared with the Neanderthal genome for several Late Pleistocene Neanderthals sequenced to low-coverage[7,13] is between 69% and 75% (Fig. 2). Thus, DNA sequences of the SH hominins diverged more than twice as far back along the lineage from the Altai Neanderthal genome to its ancestor shared with the Denisovan genome than DNA sequences of the Late Pleistocene Neanderthals from Europe and the Caucasus.

Because nearly 30 hominin skeletons have been found in SH, it is likely that the specimens analysed here belong to different individuals. The nuclear DNA sequences of femur AT-5431 and the incisor show that they belonged to the Neanderthal evolutionary lineage, and the limited data available for the molar suggest that the same is true for this specimen. Thus, the results show that the SH hominins were early Neanderthals or closely related to the ancestors of Neanderthals after the divergence from a common ancestor shared with Denisovans. Although it is difficult to determine the age of Middle Pleistocene sites with certainty, geological dating methods[1], as well as the length of the branches in trees relating the mtDNAs from femur XIII and an SH cave bear to other mtDNAs[2,12], suggest an age of around 400,000 years for the SH fossils. This age is compatible with the population split time of 381,000–473,000 years ago estimated for Neanderthals and Denisovans on the basis of their nuclear genome sequences and using the human mutation rate of $0.5 \times 10^{-9}$ per base pair per year[7]. This mutation rate also suggests that the population split between archaic and modern humans occurred between 550,000 and 765,000 years ago. Such an ancient separation of archaic and modern humans is difficult to reconcile with the suggestion that younger specimens often classified as *Homo heidelbergensis*, for example Arago or Petralona, belong to a population ancestral both to modern humans and to Neanderthals[15].

We further note that the SH hominins carry mtDNAs more closely related to those of Denisovans in Asia than Neanderthals, even though their nuclear genomes show that they are more closely related to Neanderthals. We have previously speculated that this discrepancy may be because the SH hominins carried two very divergent mtDNA lineages or that another hominin group contributed mtDNA both to the SH hominins and to Denisovans[2]. However, given that the SH hominins are early Neanderthals (or closely related to these), and assuming that the mtDNA they carried was typical of early Neanderthals, an additional possibility that appears reasonable is that the mtDNAs seen in Late Pleistocene Neanderthals were acquired by them later, presumably because of gene flow from Africa. It is possible that contacts between Africa and western Eurasia occurred in the Middle Pleistocene as indicated, for example, by the appearance of the Acheulean hand axe technology in Eurasia by 500,000 years ago[16] and by the spread of the so-called 'Mode 3' technology around 250,000 years ago[17]. Gene flow from Africa may perhaps also explain the absence of Neanderthal-derived morphological traits in some Middle Pleistocene specimens in Europe such as Ceprano and Mala Balanica[18,19]. Retrieval of further mtDNAs and, if possible, nuclear DNA from Middle Pleistocene fossils will be necessary to comprehensively address how Middle and Late Pleistocene hominins in Eurasia were related to each other.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 16 September 2015; accepted 2 February 2016.**
**Published online 14 March 2016.**

1. Arsuaga, J. L. *et al.* Neandertal roots: cranial and chronological evidence from Sima de los Huesos. *Science* **344,** 1358–1363 (2014).
2. Meyer, M. *et al.* A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* **505,** 403–406 (2014).

3. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468,** 1053–1060 (2010).
4. Sawyer, S. *et al.* Nuclear and mitochondrial DNA sequences from two Denisovan individuals. *Proc. Natl Acad. Sci. USA* **112,** 15696–15700 (2015).
5. Qin, P. & Stoneking, M. Denisovan ancestry in East Eurasian and Native American populations. *Mol. Biol. Evol.* **32,** 2665–2674 (2015).
6. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338,** 222–226 (2012).
7. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505,** 43–49 (2014).
8. Arsuaga, J. L. *et al.* Postcranial morphology of the middle Pleistocene humans from Sima de los Huesos, Spain. *Proc. Natl Acad. Sci. USA* **112,** 11524–11529 (2015).
9. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl Acad. Sci. USA* **104,** 14616–14621 (2007).
10. Krause, J. *et al.* The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature* **464,** 894–897 (2010).
11. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Pääbo, S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS ONE* **7,** e34131 (2012).
12. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl Acad. Sci. USA* **110,** 15758–15763 (2013).
13. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328,** 710–722 (2010).
14. Rasmussen, M. *et al.* An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science* **334,** 94–98 (2011).
15. Stringer, C. The status of *Homo heidelbergensis* (Schoetensack 1908). *Evol. Anthropol.* **21,** 101–107 (2012).
16. Lycett, S. J. Understanding ancient hominin dispersals using artefactual data: a phylogeographic analysis of Acheulean handaxes. *PLoS ONE* **4,** e7404 (2009).
17. Lahr, M. M. & Foley, R. A. Towards a theory of modern human origins: Geography, demography, and diversity in recent human evolution. *Yb. Phys. Anthropol.* **41,** 137–176 (1998).
18. Dennell, R. W., Martinon-Torres, Bermúdez de Castro, J. M. Hominin variability, climatic instability and population demography in Middle Pleistocene Europe. *Quat. Sci. Rev.* **30,** 1511–1524 (2011).
19. Rink, W. J. *et al.* New radiometric ages for the BH-1 hominin from Balanica (Serbia): implications for understanding the role of the Balkans in Middle Pleistocene human evolution. *PLoS ONE* **8,** e54608 (2013).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank B. Höber and A. Weihmann for help with sequencing the libraries, G. Renaud for processing the raw sequence data, S. Castellano and U. Stenzel for discussions and comments on the manuscript. Genetics work was funded by the Max Planck Society and its Presidential Innovation Fund. Field work at the Sierra de Atapuerca sites was funded by the Junta de Castilla y Leon, the Fundacion Atapuerca, the Spanish Ministerio de Ciencia e Innovacion (project CGL2009-12703-C03) and the Spanish Ministerio de Economia y Competitividad (project CGL2012-38434-C03).

**Author Contributions** M.M., J.-L.A. and S.P. directed the experimental work and wrote the manuscript. M.M. designed the laboratory experiments, which S.N., A.A. and B.N. performed. M.M., C.d.F., B.V., J.K. and K.P. analysed the data. J.-L.A., I.M., A.G., J.M.B. and E.C. excavated the fossil and provided archaeological expertise.

**Author Information** Sequences generated in this study have been deposited in the European Nucleotide Archive under study accession number PRJEB10597. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.M. (mmeyer@eva.mpg.de).

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

**Sampling, DNA extraction and library preparation.** Over the past decade, several specimens from the SH excavation were exempted from the routine cleaning procedures for molecular analysis. Samples were removed from four of these specimens in a laboratory dedicated to ancient DNA work: (1) molar AT-5444, excavated in 2006 in Square T-16, Level LU-6, was sampled by drilling into one of the roots with a dentistry drill; (2) incisor AT-5482, removed from Square U-17, Level LU-6 in 2010, from which the tip of the root was removed with a scalpel, and powder was removed by drilling into the dentine; (3) scapula AT-6672, excavated in 2012 in Square U-16, Level LU-6, was sampled by drilling powder from the specimen; (4) femur fragment AT-5431, excavated in 2014 from Square T-12, Level LU-6, was sampled by scratching off thin slices of bone from the wall of the marrow cavity with a scalpel. Between 9 and 94 mg of material sample was used for DNA extraction by a silica-based method optimized for the recovery of short DNA fragments[12]. In an attempt to remove parts of the microbial DNA contamination, the sample from femur AT-5431 was treated with 0.5 M sodium phosphate buffer (pH 7.0) before DNA extraction[20]. Between 30% and 60% of the extracts were converted to DNA libraries using single-stranded library preparation[21]; in some cases an optimized version of the original method[20] was used (see Extended Data Table 1). The number of molecules in each library was determined by digital droplet PCR[22] and libraries were barcoded through amplification with pairs of indexed primers[23] using AccuPrime Pfx DNA polymerase (Life Technologies)[24]. Extraction and library negative controls were carried through DNA extraction, library preparation, mtDNA enrichment and sequenced alongside the SH samples to monitor laboratory contamination.

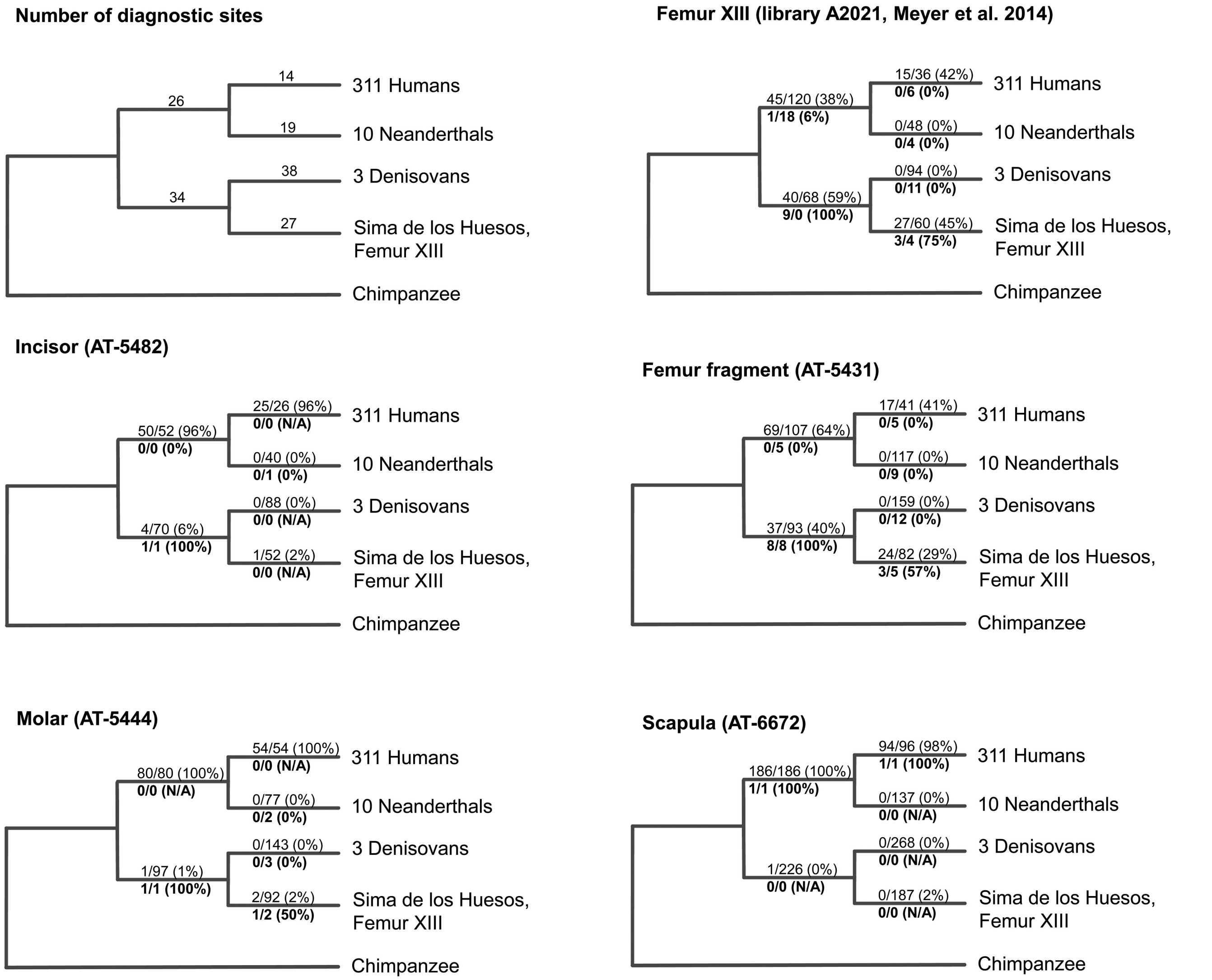
**Mitochondrial DNA enrichment and analysis.** Enrichment of mtDNA was performed in single or two successive rounds of hybridization capture following ref. 25 but using human mtDNA baits recovered from a microarray[26] and using reduced temperatures in the hybridization and wash steps as described elsewhere[2]. Enriched libraries were pooled with other libraries and sequenced on Illumina MiSeq and HiSeq 2500 platforms using 76 bp paired-end sequencing recipes for double-indexed sequencing[23]. Base calling was performed using Illumina Bustard software (MiSeq) or FreeIbis[27] (HiSeq). Sequences that did not perfectly match one of the expected index combinations were discarded. Overlapping paired-end reads were merged into single sequences to reconstruct full-length molecules[28]. Merged sequences $\geq$30 bp were aligned to the revised Cambridge reference human mtDNA sequence (rCRS; accession number NC_012920) using BWA[29] with ‘ancient’ parameters[6] requiring a mapping quality score $\geq$30. Duplicate sequences were removed by calling a consensus from sequences with identical alignment start and end coordinates (bam-rmdup; https://github.com/udo-stenzel/biohazard). Putatively deaminated fragments were identified on the basis of the presence of a C to T substitution to the reference genome in the first or last position.

To identify positions in the mtDNA genome that were diagnostic for each branch of the hominin tree, we aligned the mtDNA sequences from 311 worldwide present-day humans[30], 10 Neanderthals[7,31–33], 3 Denisovans[3,4], SH femur XIII[2] and the chimpanzee[34] to the rCRS using MAFFT[35]. To exclude errors due to cytosine deamination when counting the number of sequences supporting the respective branch, we used the fact that single-stranded library preparation is strand-specific, and we disregarded all sequences aligned to the reference genome in the orientation as sequenced if one of the two states of a diagnostic site was C, and sequences aligned in the reverse complement direction if one of the two states was G.

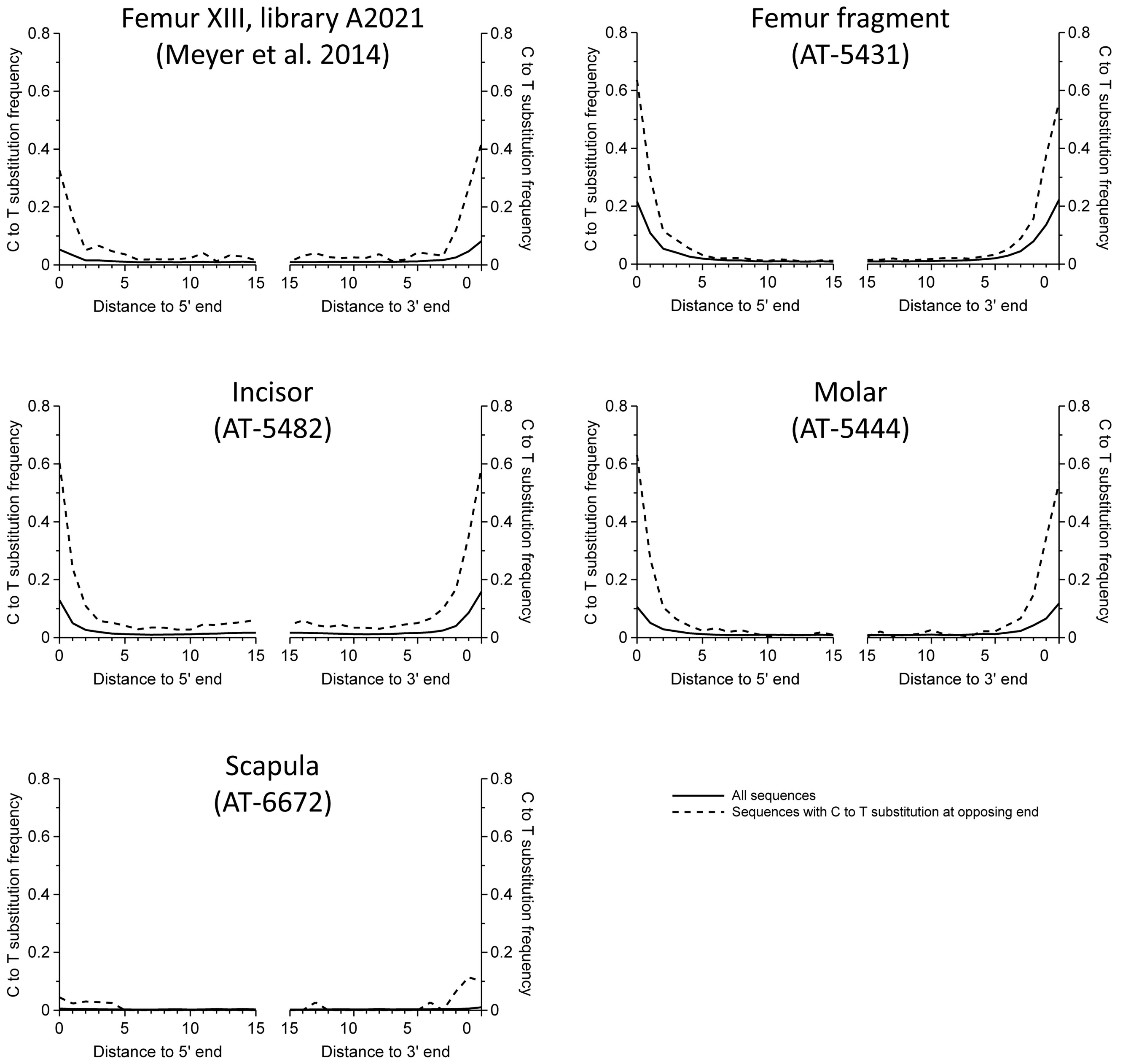
**Nuclear DNA sequencing and raw data processing.** In an initial attempt to obtain nuclear DNA sequences from an SH hominin, we sequenced eight lanes of a previously prepared library (library A2021) from femur XIII[2] on an Illumina HiSeq 2500 instrument using the same recipe as for the libraries enriched for mtDNA, but with shorter read length (2 × 50 bp). A 0.5% spike-in of a double-indexed $\varphi$X174 control library (P7 index CGATTCG, P5 index CGATTCG) was added to the library to allow base calling with FreeIbis. The absence of sequence complexity in the index reads forced us to drop index filtering for this run. We then generated data from the other four specimens using between four and seven HiSeq lanes each (2 × 50 or 2 × 76 bp). We spiked-in between 1.5 and 20% of a pool of four $\varphi$X174 libraries carrying AAAAAAA, CCCCCCC, GGGGGGG and TTTTTTT as P7 and P5 indices. This recovered the quality of the index reads and allowed us to filter for a perfect match to at least one of the two indices without losing large amounts of data (Extended Data Table 5). Further processing and mapping were performed as described for the capture-enriched libraries above, using only sequences of at least 35 bp for mapping to the human genome (hg19/GRCh37) and requiring a map quality score of 30 or greater.

**Phylogenetic analysis.** To place the SH specimens in the hominin phylogenetic tree, we identified phylogenetically informative positions where a randomly drawn allele from the Altai Neanderthal, Denisovan and/or Mbuti VCF files differed from all great apes (the chimpanzee (pantro2), the bonobo (panpan1), the gorilla (gorGor3), the orangutan (ponAbe2)) and the rhesus macaque (rheMac2). Orthologous ape and monkey outgroup sequences were extracted from pairwise lastz alignments to the human reference genome (hg19/GRCh37) provided by the University of California, Santa Cruz (UCSC) genome browser and in-house, and were required to show the identical base at informative positions. All sites were required to pass the map35_100% criteria described in supplementary section 5b of ref. 7. A file of all phylogenetically informative positions is available for download at http://bioinf.eva.mpg.de/sima/. To reduce the impact of cytosine deamination, all thymines within the first and last three positions of each sequence were masked out. This filter is less strict than that based on alignment orientation described above and thus retains more data. We also included sequence data of an early modern human[36], several Neanderthals[7,13] and two Denisovans[4] for comparison (Extended Data Fig. 4 and Extended Data Table 3).

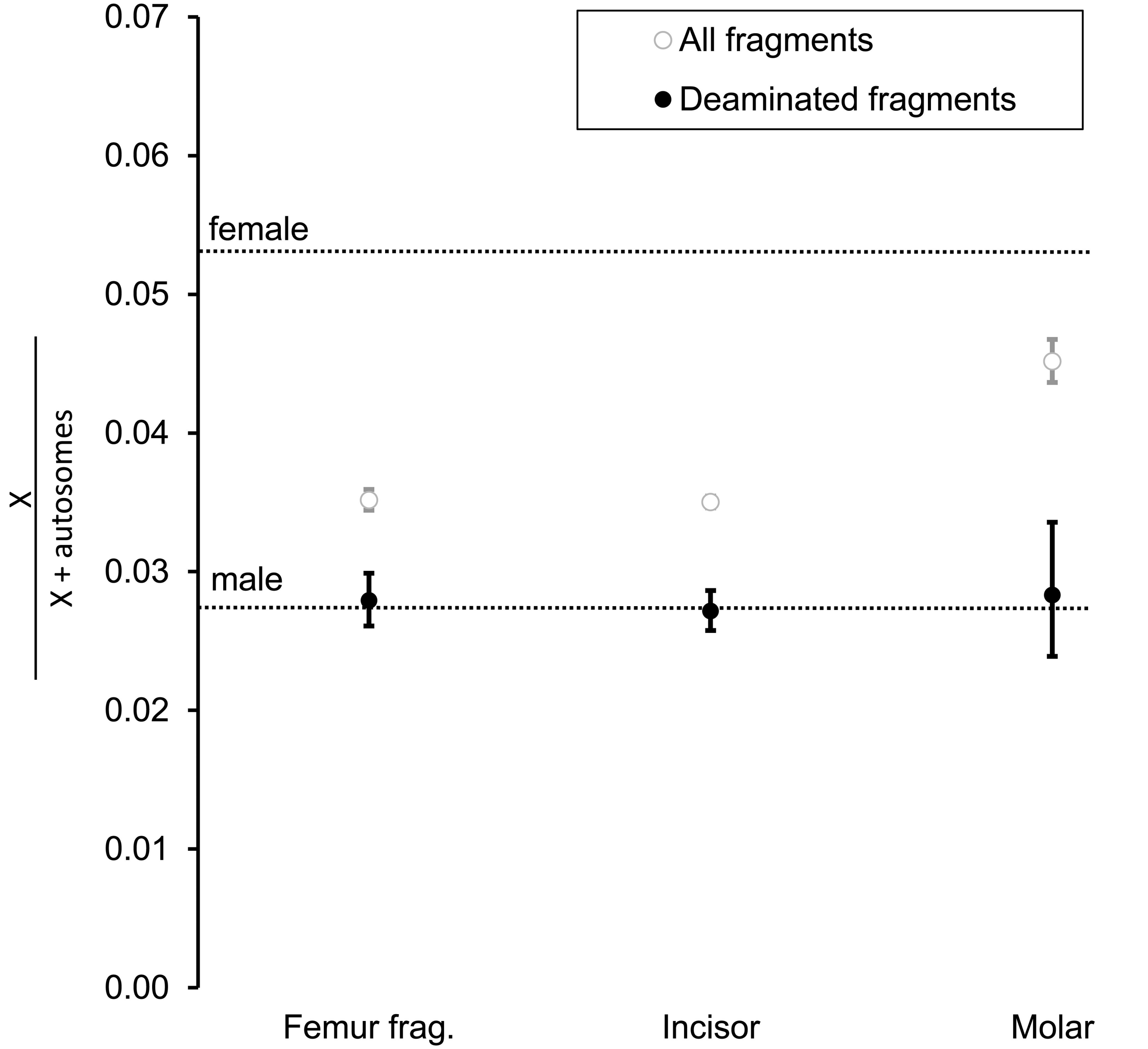
20. Korlevic´, P. *et al.* Reducing microbial and human contamination in DNA extractions from ancient bones and teeth. *Biotechniques* **59,** 87–93 (2015).
21. Gansauge, M. T. & Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature Protocols* **8,** 737–748 (2013).
22. Slon, V., Glocke, I., Barkai, R., Gopher, A., Hershkovitz, I. & Meyer, M. Mammalian mitochondrial capture, a tool for rapid screening of DNA preservation in faunal and undiagnostic remains, and its application to Middle Pleistocene specimens from Qesem Cave (Israel). *Quat. Int.* http://dx.doi.org/10.1016/j.quaint.2015.03.039 (2015).
23. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40,** e3 (2012).
24. Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52,** 87–94 (2012).
25. Maricic, T., Whitten, M. & Pääbo, S. Multiplexed DNA sequence capture of mitochondrial genomes using PCR products. *PLoS ONE* **5,** e14004 (2010).
26. Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl Acad. Sci. USA* **110,** 2223–2227 (2013).
27. Renaud, G., Kircher, M., Stenzel, U. & Kelso, J. freeIbis: an efficient basecaller with calibrated quality scores for Illumina sequencers. *Bioinformatics* **29,** 1208–1209 (2013).
28. Renaud, G., Stenzel, U. & Kelso, J. leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Res.* **42,** e141 (2014).
29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
30. Green, R. E. *et al.* A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134,** 416–426 (2008).
31. Briggs, A. W. *et al.* Targeted retrieval and analysis of five Neandertal mtDNA genomes. *Science* **325,** 318–321 (2009).
32. Gansauge, M. T. & Meyer, M. Selective enrichment of damaged DNA molecules for ancient genome sequencing. *Genome Res.* **24,** 1543–1549 (2014).
33. Skoglund, P. *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl Acad. Sci. USA* **111,** 2229–2234 (2014).
34. Horai, S. *et al.* Man's place in Hominoidea revealed by mitochondrial DNA genealogy. *J. Mol. Evol.* **35,** 32–43 (1992).
35. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
36. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–449 (2014).



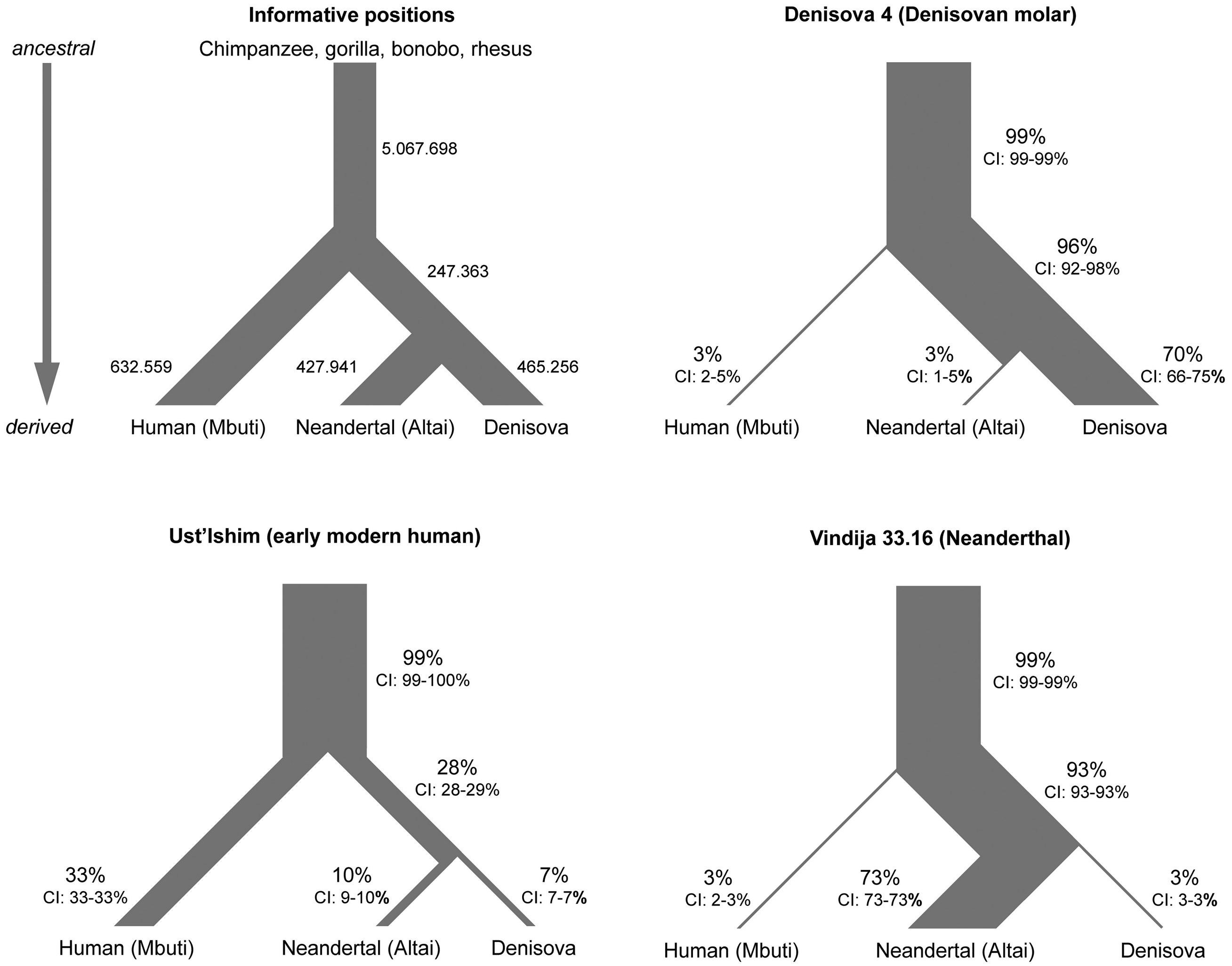
**Extended Data Figure 1 | Sharing of derived alleles at diagnostic positions separating the hominin groups in the mitochondrial tree.** The chimpanzee was used as outgroup to determine the ancestral state, which is shared with all individuals in the tree except those belonging to the labelled branch. Provided are the number of diagnostic sites available for this analysis (top left panel) as well as the number of sequences supporting the derived state, their percentage (in brackets) and the total number of observations. Numbers above the branch include all sequences whereas bold numbers below the branch are limited to sequences showing evidence of cytosine deamination. Published data from library A2021 of femur XIII were included in this analysis for comparison.



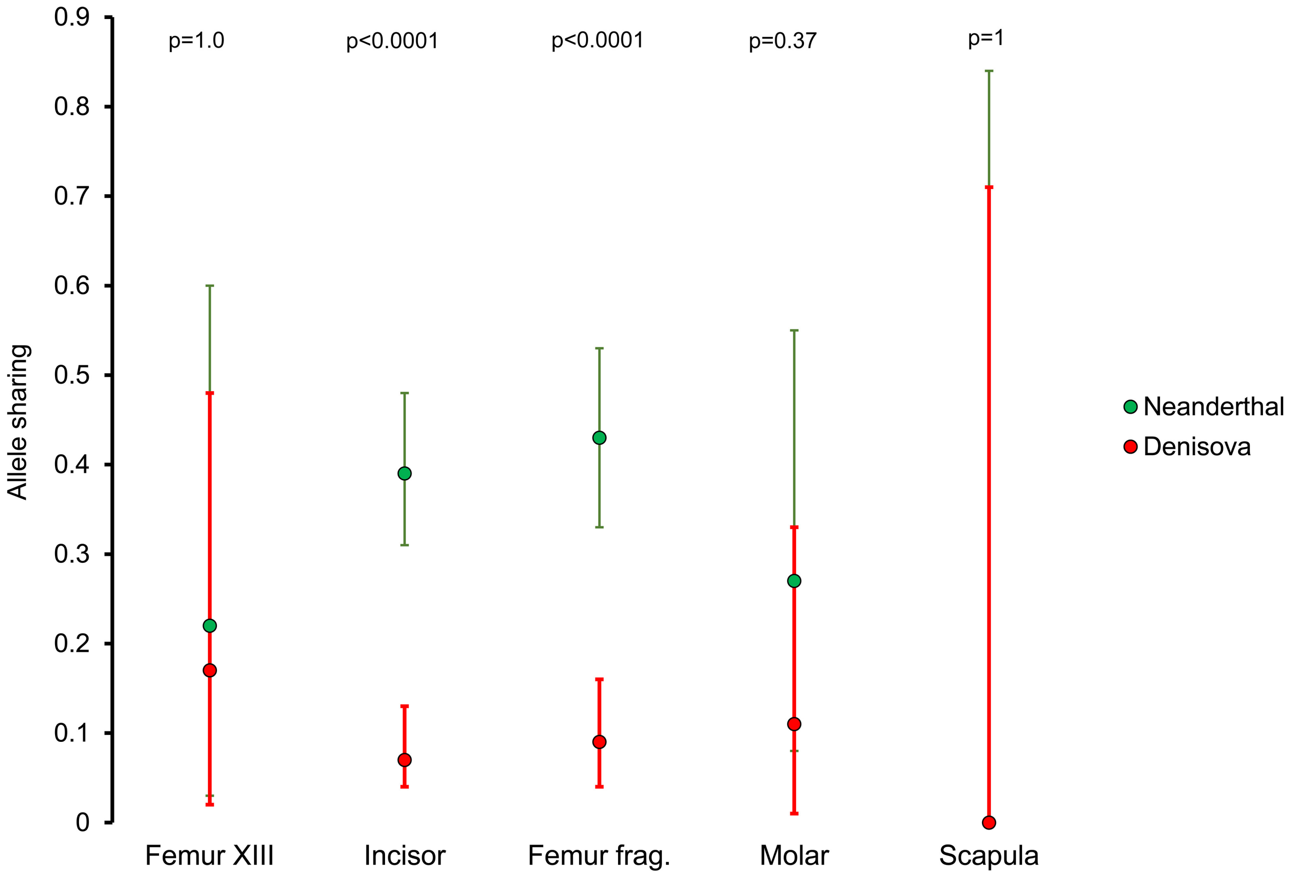
**Extended Data Figure 2 | Frequency of C to T substitutions at the beginning and end of nuclear sequence alignments.** Solid lines denote all sequences and dashed lines only those sequences carrying a C to T substitution at the opposing end.



**Extended Data Figure 3 | Sex determination based on the number of sequences aligning to chromosome X and the autosomes.** Ninety-five per cent binomial confidence intervals are provided as well as the expected ratios of X to (X + autosomal) sequences for male and female samples. The analysis was performed with and without enrichment of endogenous DNA by filtering for the presence of C to T substitutions at terminal alignment positions. Present-day human contamination in the unfiltered sequences appears to have been introduced at least partly by female individuals.



**Extended Data Figure 4 | Number of informative positions identified for each branch of the tree.** For comparison with the SH sequence data, we show the sharing of derived alleles at these positions using published sequence data from a Neanderthal (Vindija 33.16), a Denisovan individual (Denisova 4) and an early modern human (Ust'Ishim).



**Extended Data Figure 5 | Derived allele sharing with the Neanderthal- and Denisovan-specific branches in deaminated DNA fragments from all five specimens from SH.** Only sequences with a terminal C to T substitution were used in this analysis. Error bars, 95% confidence intervals. Significance was tested using Fisher's exact test (two-tailed).

**Extended Data Table 1 | Overview of DNA extracts, libraries and shotgun sequences generated in three experiments**

| Exp. | Specimen | Material used [mg] | Fraction of extract used in library prep. | Library prep. protocol | Library ID | P7 / P5 index seq. | #Molecules in library acc. to dPCR/ qPCR(*) | #seq. generated | #seq. ≥35bp | #mapped seq. | #unique seq. | Deaminated sequences (unique) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | # | Av. size |
| (Meyer et al. 2014) | Femur XIII (AT2944) | 63 | 0.6 | Ref. 21 | A2021 | ATGAGCA / CGACGGT | 5.73E+09 (*) | 2.55E+09 | 7.27E+08 | 1.52E+05 | 1.36E+05 | 2,772 | 41.7 |
| 1 | Incisor (AT-5482) | 9 | 0.3 | Ref. 21 | B2523 | CGTAATC / AATAGTA | 9.67E+08 (*) | 8.73E+08 | 4.42E+08 | 1.10E+06 | 7.27E+05 | 30,950 | 40.5 |
| 1 | Extraction blank | | 0.3 | Ref. 21 | B2524 | CTATACG / CTGCGAC | 3.78E+08 (*) | | | | | | |
| 1 | Extraction blank | | 0.3 | Ref. 21 | B2529 | TGGCGCT / TGGCAAT | 4.23E+08 (*) | | | | | | |
| 1 | Library blank | | | Ref. 21 | B2531 | ATCGTTC / CGAGATC | 2.53E+08 (*) | | | | | | |
| 2 | Femur frag. (AT-5431) | 94 | 0.3 | Ref. 20 | R1848 | AGACTCC / CCTAGGT | 1.60E+09 | 7.98E+08 | 3.53E+08 | 3.08E+05 | 2.35E+05 | 17,265 | 40.5 |
| 2 | Extraction blank | | 0.3 | Ref. 20 | R1858 | CGCTATT / ACTATCA | 4.44E+07 | | | | | | |
| 2 | Library blank | | | Ref. 20 | R1861 | GACCGAT / TAATGCG | 2.13E+07 | | | | | | |
| 3 | Molar (AT-5444) | 28 | 0.3 | Ref. 20 | R1753 | TACTCGC / CTCGATG | 2.34E+09 | 8.93E+08 | 4.89E+08 | 8.05E+04 | 6.95E+04 | 2,635 | 40.5 |
| 3 | Scapula (AT-6672) | 20 | 0.3 | Ref. 20 | R1754 | AGCGCCA / GCTCGAA | 1.14E+10 | 5.76E+08 | 3.49E+08 | 9.99E+04 | 9.75E+04 | 274 | 60.6 |
| 3 | Extraction blank | | 0.3 | Ref. 20 | R1757 | GTTGCAT / AACTCCG | 2.45E+07 | | | | | | |
| 3 | Library blank | | | Ref. 20 | R1758 | ATCCTCT / TTGAAGT | 9.15E+06 | | | | | | |

Reported are the amount of sample used for DNA extraction, the fraction of DNA extract converted into DNA library, the number of molecules in each library as determined by digital PCR (dPCR) or quantitative PCR (qPCR), the protocol used for library preparation, index sequences of each library, the number of sequences generated from each library and the number of sequences retained after length filtering, mapping, duplicate removal and identification of putatively deaminated fragments.

**Extended Data Table 2 | Characteristics of sequences obtained after mtDNA enrichment**

| Exp. | Specimen | Library ID | #sequences | #overlap-merged sequences ≥30bp | #mapped sequences (MQ ≥30) | #unique seq. | Dup. rate | #Seq. with C to T substitution in terminal positions | C to T substitution frequency [%] (#observations) | | Human contamination [%] (95% C.I.) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | 5' end | 3' end | All fragments | Deaminated fragments |
| (Meyer et al. 2014) | Femur XIII (AT2944) | A2021 | | | 257,277 | 2,592 | 97.3 | 687 | 46.4<br>(250/539) | 47.8<br>(249/521) | 41.7<br>(25.5-59.2) | 0.0<br>(0.0-45.9) |
| 1 | Incisor (AT-5482) | B2523 | 1,172,042 | 773,601 | 62,659 | 1,419 | 44.0 | 47 | 4.7<br>(14/298) | 6.7<br>(14/208) | 96.2<br>(80.4-99.9) | N/A |
| 1 | Extraction blank | B2524 | 74,304 | 44,231 | 9,792 | 151 | 64.3 | 3 | 0.0<br>(0/30) | 0.0<br>(0/29) | | |
| 1 | Extraction blank | B2529 | 61,439 | 42,600 | 10,009 | 189 | 52.7 | 0 | 0.0<br>(0/47) | 0.0<br>(0/39) | | |
| 1 | Library blank | B2531 | 56,078 | 31,179 | 8,827 | 71 | 124.0 | 0 | 0.0<br>(0/15) | 0.0<br>(0/16) | | |
| 2 | Femur frag. (AT-5431) | R1848 | 38,797,693 | 25,567,161 | 678,518 | 3,188 | 207.7 | 731 | 44.2<br>(277/627) | 41.1<br>(233/567) | 41.5<br>(26.3-57.9) | 0.0<br>(0.0-52.2) |
| 2 | Extraction blank | R1858 | 1,096,093 | 562,893 | 201,219 | 1,361 | 147.9 | 16 | 1.5<br>(4/261) | 1.3<br>(3/229) | | |
| 2 | Library blank | R1861 | 125,743 | 43,344 | 6,377 | 101 | 63.1 | 2 | 6.3<br>(1/16) | 0<br>(0/15) | | |
| 3 | Molar (AT-5444) | R1753 | 43,834,971 | 34,002,764 | 105,628 | 3,041 | 34.5 | 52 | 3.0<br>(15/498) | 3.0<br>(20/666) | 100.0<br>(93.4-100.0) | N/A |
| 3 | Scapula (AT-6672) | R1754 | 34,645,495 | 27,006,518 | 49,676 | 3,742 | 13.2 | 31 | 0.5<br>(4/784) | 1.0<br>(7/721) | 97.9<br>(92.7-99.7) | 100.0<br>(2.5-100.0) |
| 3 | Extraction blank | R1757 | 296,666 | 103,470 | 7,608 | 358 | 21.1 | 3 | 1.5<br>(1/67) | 0<br>(0/76) | | |
| 3 | Library blank | R1758 | 565,241 | 156,407 | 10,261 | 58 | 166.7 | 3 | 0 (0/5) | 0 (0/6) | | |

Extracts and sample libraries and their related controls were generated in three independent experiments as indicated. Provided are the total number of sequences generated from each library, the number of sequences ≥30 bp, the number of sequences ≥30 bp that could be mapped to the revised Cambridge reference sequence, the number of sequences remaining after duplicate removal and the average number of sequences collapsed into unique sequences during duplicate removal. Putatively deaminated fragments were identified by C to T substitutions at the terminal alignment positions, the frequency of which is denoted. Estimates of present-day human contamination and the respective 95% binomial confidence intervals were determined on the basis of the percentage of sharing of human-derived alleles (compare Extended Data Fig. 1). Previously published data from femur XIII (library A2021) are included for comparison.

**Extended Data Table 3 | Fraction of derived alleles shared with the human, Neanderthal and Denisovan genomes and combinations thereof**

a

| | Included sequences | Human | Neanderthal | Denisova | Neanderthal-Denisova | Human-Neanderthal-Denisova | Human-Denisova | Neanderthal-Human |
|---|---|---|---|---|---|---|---|---|
| Femur XIII (AT-2944) | deaminated | 0.23 (0.08-0.45) | 0.22 (0.03-0.60) | 0.17 (0.02-0.48) | 0.67 (0.22-0.96) | 0.98 (0.94-0.99) | 0.20 (0.01-0.72) | 1.00 (0.40-1.00) |
| | all | 0.31 (0.29-0.33) | 0.13 (0.11-0.15) | 0.09 (0.07-0.10) | 0.30 (0.27-0.34) | 0.99 (0.98-0.99) | 0.64 (0.58-0.70) | 0.70 (0.65-0.76) |
| Incisor (AT-5482) | deaminated | 0.05 (0.02-0.09) | 0.39 (0.31-0.48) | 0.07 (0.04-0.13) | 0.68 (0.55-0.80) | 0.99 (0.98-0.99) | 0.29 (0.13-0.51) | 0.82 (0.62-0.94) |
| | all | 0.31 (0.30-0.32) | 0.13 (0.12-0.14) | 0.07 (0.06-0.07) | 0.34 (0.33-0.36) | 0.98 (0.98-0.99) | 0.64 (0.61-0.67) | 0.69 (0.67-0.72) |
| Femur frag. (AT-5431) | deaminated | 0.05 (0.02-0.09) | 0.43 (0.33-0.53) | 0.09 (0.04-0.16) | 0.87 (0.75-0.95) | 1.00 (0.99-1.00) | 0.39 (0.17-0.64) | 0.67 (0.47-0.83) |
| | all | 0.25 (0.23-0.27) | 0.17 (0.15-0.19) | 0.07 (0.06-0.09) | 0.44 (0.41-0.47) | 0.99 (0.99-0.99) | 0.55 (0.50-0.60) | 0.70 (0.66-0.74) |
| Molar (AT-5444) | deaminated | 0.00 (0.00-0.23) | 0.27 (0.08-0.55) | 0.11 (0.01-0.33) | 0.63 (0.25-0.92) | 0.99 (0.97-1.00) | 0.33 (0.01-0.91) | 1.00 (0.03-1.00) |
| | all | 0.29 (0.25-0.32) | 0.12 (0.10-0.16) | 0.08 (0.06-0.11) | 0.33 (0.28-0.38) | 0.99 (0.98-0.99) | 0.61 (0.52-0.70) | 0.67 (0.58-0.76) |
| Scapula (AT-6672) | deaminated | 1.00 (0.03-1.00) | 0.00 (0.00-0.84) | 0.00 (0.00-0.71) | 0.67 (0.09-0.99) | 1.00 (0.83-1.00) | 0.00 (0.00-0.98) | N/A |
| | all | 0.32 (0.30-0.35) | 0.11 (0.09-0.13) | 0.06 (0.05-0.07) | 0.29 (0.26-0.33) | 0.99 (0.98-0.99) | 0.62 (0.56-0.68) | 0.71 (0.66-0.76) |
| Denisova 4 | deaminated | 0.03 (0.02-0.05) | 0.03 (0.01-0.05) | 0.70 (0.66-0.75) | 0.96 (0.92-0.98) | 0.99 (0.99-0.99) | 0.86 (0.76-0.92) | 0.12 (0.06-0.20) |
| Denisova 8 | deaminated | 0.02 (0.01-0.02) | 0.06 (0.06-0.06) | 0.59 (0.58-0.60) | 0.92 (0.91-0.93) | 1.00 (0.99-1.00) | 0.88 (0.87-0.90) | 0.14 (0.13-0.16) |
| Feldhofer 1 | deaminated | 0.05 (0.03-0.08) | 0.69 (0.63-0.75) | 0.03 (0.01-0.06) | 0.93 (0.87-0.96) | 0.99 (0.99-1.00) | 0.09 (0.02-0.24) | 0.86 (0.75-0.93) |
| Feldhofer 2 | deaminated | 0.00 (0.00-0.15) | 0.75 (0.48-0.93) | 0.06 (0.00-0.30) | 0.86 (0.42-1.00) | 0.99 (0.97-1.00) | 0.33 (0.01-0.91) | 1.00 (0.63-1.00) |
| Mezmaiskaya 1 | deaminated | 0.03 (0.03-0.03) | 0.74 (0.73-0.74) | 0.03 (0.03-0.03) | 0.92 (0.92-0.92) | 1.00 (1.00-1.00) | 0.11 (0.11-0.12) | 0.91 (0.91-0.91) |
| Vindija 33.16 | deaminated | 0.03 (0.02-0.03) | 0.73 (0.73-0.73) | 0.03 (0.03-0.03) | 0.93 (0.93-0.93) | 0.99 (0.99-0.99) | 0.11 (0.11-0.11) | 0.91 (0.91-0.92) |
| Vindija 33.25 | deaminated | 0.02 (0.02-0.02) | 0.75 (0.74-0.75) | 0.03 (0.03-0.03) | 0.94 (0.93-0.94) | 0.99 (0.99-0.99) | 0.10 (0.10-0.10) | 0.91 (0.91-0.92) |
| Vindija 33.26 | deaminated | 0.02 (0.02-0.02) | 0.74 (0.74-0.74) | 0.03 (0.03-0.03) | 0.93 (0.93-0.94) | 0.99 (0.99-0.99) | 0.10 (0.10-0.10) | 0.91 (0.91-0.92) |
| Ust'Ishim | deaminated | 0.33 (0.33-0.33) | 0.10 (0.09-0.10) | 0.07 (0.07-0.07) | 0.28 (0.28-0.29) | 0.99 (0.99-0.99) | 0.66 (0.66-0.67) | 0.71 (0.71-0.71) |

b

| | Included sequences | Human | Neanderthal | Denisova | Neanderthal-Denisova | Human-Neanderthal-Denisova | Human-Denisova | Neanderthal-Human |
|---|---|---|---|---|---|---|---|---|
| Femur XIII (AT-2944) | deaminated | 5/22 | 2/9 | 2/12 | 4/6 | 166/170 | 1/5 | 4/4 |
| | all | 529/1698 | 157/1220 | 111/1278 | 202/666 | 13680/13878 | 161/250 | 212/301 |
| Incisor (AT-5482) | deaminated | 10/199 | 51/131 | 10/137 | 41/60 | 1592/1617 | 7/24 | 22/27 |
| | all | 2725/8657 | 742/5763 | 412/6108 | 1134/3313 | 66296/67364 | 868/1357 | 1041/1498 |
| Femur frag. (AT-5431) | deaminated | 7/150 | 42/98 | 10/110 | 47/54 | 1278/1284 | 7/18 | 20/30 |
| | all | 651/2592 | 287/1666 | 143/1941 | 448/1012 | 20271/20522 | 211/386 | 325/464 |
| Molar (AT-5444) | deaminated | 0/14 | 4/15 | 2/19 | 5/8 | 165/166 | 1/3 | 1/1 |
| | all | 213/744 | 65/523 | 42/532 | 110/333 | 6081/6165 | 77/126 | 83/123 |
| Scapula (AT-6672) | deaminated | 1/1 | 0/2 | 0/3 | 2/3 | 20/20 | 0/1 | 0/0 |
| | all | 552/1699 | 127/1182 | 70/1198 | 198/682 | 13075/13256 | 166/268 | 214/301 |
| Denisova 4 | deaminated | 20/580 | 12/460 | 320/454 | 230/240 | 4958/5000 | 72/84 | 12/98 |
| Denisova 8 | deaminated | 262/15564 | 646/10780 | 6646/11274 | 5270/5720 | 127696/128310 | 2168/2456 | 380/2670 |
| Feldhofer 1 | deaminated | 17/357 | 149/215 | 6/233 | 141/152 | 2799/2820 | 3/34 | 55/64 |
| Feldhofer 2 | deaminated | 0/22 | 12/16 | 1/16 | 6/7 | 194/195 | 1/3 | 8/8 |
| Mezmaiskaya 1 | deaminated | 9183/308630 | 154251/209649 | 7039/229075 | 111722/121511 | 2489810/2499964 | 5382/47214 | 50559/55532 |
| Vindija 33.16 | deaminated | 5195/204696 | 97859/133612 | 4833/154912 | 72346/77903 | 1703087/1716488 | 3345/30304 | 33475/36604 |
| Vindija 33.25 | deaminated | 4633/207608 | 101934/136687 | 4842/155443 | 74188/79239 | 1693259/1705720 | 3156/31131 | 33449/36574 |
| Vindija 33.26 | deaminated | 4312/186298 | 91088/123221 | 4348/140193 | 66384/71122 | 1536122/1547680 | 2805/28107 | 30311/33196 |
| Ust'Ishim | deaminated | 84958/256372 | 16567/173398 | 12857/188709 | 28536/100640 | 2030409/2059155 | 26116/39296 | 31792/44800 |

**a**, Point estimates as well as 95% binomial confidence intervals. For the SH hominins, numbers are provided for all fragments ≥35 bp and the subset of fragments showing evidence of deamination.
**b**, Number of DNA fragments matching the derived state as well the total number of informative fragments.

**Extended Data Table 4 | Derived allele sharing between putatively deaminated DNA fragments of the five SH specimens and all branches of the hominin evolutionary tree**

| | | Present-day human used in analysis | | | | | |
|---|---|---|---|---|---|---|---|
| **Branch** | **Specimen** | **Mbuti** | **Yoruba** | **San** | **French** | **Han** | **Papuan** |
| Neandertal | FemurXIII | 22.2% (2/9) | 53.8% (7/13) | 46.2% (6/13) | 36.4% (4/11) | 41.7% (5/12) | 41.7% (5/12) |
| | Femur AT-5431 | 42.9% (42/98) | 39.8% (39/98) | 41.7% (40/96) | 41.0% (41/100) | 40.2% (39/97) | 44.8% (43/96) |
| | Incisor | 38.9% (51/131) | 38.9% (49/126) | 39.2% (51/130) | 39.8% (51/128) | 38.5% (47/122) | 37.2% (45/121) |
| | Molar | 26.7% (4/15) | 28.6% (4/14) | 28.6% (4/14) | 26.7% (4/15) | 30.8% (4/13) | 21.4% (3/14) |
| | Scapula | 0.0% (0/2) | 0.0% (0/3) | 0.0% (0/3) | 0.0% (0/3) | 0.0% (0/3) | 0.0% (0/3) |
| Denisova | FemurXIII | 16.7% (2/12) | 14.3% (2/14) | 12.5% (2/16) | 14.3% (2/14) | 6.7% (1/15) | 12.5% (2/16) |
| | Femur AT-5431 | 9.1% (10/110) | 5.9% (6/102) | 8.1% (9/111) | 6.4% (7/109) | 6.5% (7/108) | 7.2% (8/111) |
| | Incisor | 7.3% (10/137) | 9.1% (12/132) | 8.5% (12/141) | 8.1% (11/136) | 7.4% (10/136) | 8.1% (11/135) |
| | Molar | 10.5% (2/19) | 5.6% (1/18) | 5.3% (1/19) | 5.3% (1/19) | 5.6% (1/18) | 5.0% (1/20) |
| | Scapula | 0.0% (0/3) | 0.0% (0/3) | 0.0% (0/4) | 0.0% (0/4) | 0.0% (0/4) | 0.0% (0/3) |
| Human | FemurXIII | 22.7% (5/22) | 18.8% (3/16) | 15.4% (4/26) | 13.3% (2/15) | 25.0% (4/16) | 12.5% (2/16) |
| | Femur AT-5431 | 4.7% (7/150) | 5.5% (7/127) | 7.1% (11/154) | 5.4% (8/147) | 8.3% (12/144) | 5.7% (8/140) |
| | Incisor | 5.0% (10/199) | 6.1% (12/197) | 3.0% (6/198) | 5.3% (10/188) | 6.3% (12/190) | 8.0% (14/176) |
| | Molar | 0.0% (0/14) | 0.0% (0/20) | 0.0% (0/22) | 0.0% (0/20) | 0.0% (0/20) | 0.0% (0/17) |
| | Scapula | 100.0% (1/1) | 33.3% (1/3) | 100.0% (1/1) | 20.0% (1/5) | 33.3% (1/3) | 50.0% (1/2) |
| Neandertal-Denisova | FemurXIII | 66.7% (4/6) | 75.0% (3/4) | 71.4% (5/7) | 81.8% (9/11) | 72.7% (8/11) | 71.4% (5/7) |
| | Femur AT-5431 | 87.0% (47/54) | 81.8% (36/44) | 83.3% (40/48) | 85.7% (42/49) | 82.2% (37/45) | 80.0% (40/50) |
| | Incisor | 68.3% (41/60) | 73.5% (50/68) | 74.6% (53/71) | 71.9% (46/64) | 76.2% (48/63) | 71.4% (45/63) |
| | Molar | 62.5% (5/8) | 75.0% (6/8) | 77.8% (7/9) | 50.0% (4/8) | 55.6% (5/9) | 66.7% (4/6) |
| | Scapula | 66.7% (2/3) | 66.7% (2/3) | 0.0% (0/1) | 0.0% (0/1) | 50.0% (1/2) | 50.0% (1/2) |
| Human-Denisova | FemurXIII | 20.0% (1/5) | 50.0% (1/2) | 33.3% (1/3) | 33.3% (1/3) | 50.0% (1/2) | 50.0% (1/2) |
| | Femur AT-5431 | 38.9% (7/18) | 34.5% (10/29) | 31.8% (7/22) | 40.0% (10/25) | 40.0% (10/25) | 39.1% (9/23) |
| | Incisor | 29.2% (7/24) | 19.2% (5/26) | 21.7% (5/23) | 26.9% (7/26) | 23.1% (6/26) | 20.0% (6/30) |
| | Molar | 33.3% (1/3) | 50.0% (2/4) | 50.0% (2/4) | 66.7% (2/3) | 50.0% (2/4) | 100.0% (2/2) |
| | Scapula | 0.0% (0/1) | NA% (0/0) | NA% (0/0) | NA% (0/0) | NA% (0/0) | 0.0% (0/1) |
| Neandertal-Human | FemurXIII | 100.0% (4/4) | NA% (0/0) | 100.0% (1/1) | 100.0% (2/2) | 100.0% (2/2) | 100.0% (2/2) |
| | Femur AT-5431 | 66.7% (20/30) | 68.8% (22/32) | 67.7% (21/31) | 77.8% (21/27) | 77.4% (24/31) | 63.3% (19/30) |
| | Incisor | 81.5% (22/27) | 77.4% (24/31) | 70.0% (21/30) | 80.0% (24/30) | 75.8% (25/33) | 67.6% (25/37) |
| | Molar | 100.0% (1/1) | 50.0% (1/2) | 50.0% (1/2) | 100.0% (1/1) | 50.0% (1/2) | 100.0% (2/2) |
| | Scapula | NA% (0/0) | NA% (0/0) | NA% (0/0) | NA% (0/0) | NA% (0/0) | NA% (0/0) |
| All | FemurXIII | 97.6% (166/170) | 96.0% (170/177) | 97.1% (169/174) | 96.3% (157/163) | 97.0% (164/169) | 97.1% (169/174) |
| | Femur AT-5431 | 99.5% (1278/1284) | 99.5% (1299/1306) | 99.4% (1298/1306) | 99.5% (1303/1310) | 99.5% (1297/1303) | 99.5% (1299/1305) |
| | Incisor | 98.5% (1592/1617) | 98.3% (1603/1630) | 98.3% (1605/1632) | 98.4% (1621/1647) | 98.2% (1591/1620) | 98.4% (1620/1647) |
| | Molar | 99.4% (165/166) | 99.4% (168/169) | 98.8% (169/171) | 100.0% (173/173) | 100.0% (169/169) | 99.4% (172/173) |
| | Scapula | 100.0% (20/20) | 100.0% (21/21) | 100.0% (22/22) | 100.0% (23/23) | 100.0% (22/22) | 100.0% (21/21) |

Phylogenetically informative sites were identified using six present-day human genomes. Provided are the fraction of shared derived alleles and the counts in brackets.

**Extended Data Table 5 | Overview of the sequencing runs performed**

| Specimen | PhiX spike-in [%] | FlowCellID | #Sequencing cycles | Lane | One perfect index match to sample library [%] | Two perfect index matches to sample library [%] |
|---|---|---|---|---|---|---|
| Femur XIII (AT-2944), library A2021 | 0.5 | 130827_SN7001204_0226 | 2x 50 | 1 | 0.7 | 0.0 |
| | | | | 2 | 15.8 | 0.0 |
| | | | | 3 | 24.4 | 0.0 |
| | | | | 4 | 37.0 | 2.0 |
| | | | | 5 | 40.2 | 6.7 |
| | | | | 6 | 38.9 | 2.7 |
| | | | | 7 | 24.9 | 0.0 |
| | | | | 8 | 10.5 | 0.0 |
| Incisor (AT-5482) | 1.5 | 140718_SN7001204_0282 | 2x 76 | 1 | 68.2 | 8.3 |
| | 4 | 150126_SN7001204_0332 | 2x 50 | 1 | 45.6 | 5.6 |
| | | | | 2 | 79.7 | 29.1 |
| | 15 | 150127_SN7001204_0333 | 2x 76 | 2 | 85.1 | 69.2 |
| | 20 | 150212_SN7001204_0335 | 2x 76 | 2 | 85.1 | 69.2 |
| | 20 | 150212_SN7001204_0336 | 2x 50 | 1 | 60.9 | 23.5 |
| | | | | 2 | 75.9 | 35.5 |
| Femur fragment (AT-5431) | 20 | 150312_SN7001204_0345 | 2x 50 | 1 | 82.2 | 61.4 |
| | | | | 2 | 83.1 | 64.2 |
| | | | | 3 | 83.1 | 67.5 |
| | | | | 4 | 82.0 | 64.6 |
| Molar (AT-5444) | 20 | 150302_SN7001204_0343 | 2x 50 | 1 | 54.2 | 0.1 |
| | | | | 2 | 64.7 | 2.1 |
| | | 150423_SN7001204_0374 | 2x 50 | 1 | 83.7 | 66.0 |
| | | | | 2 | 85.5 | 72.7 |
| | | 150423_SN7001204_0375 | 2x 50 | 1 | 82.3 | 31.8 |
| | | | | 2 | 82.1 | 28.4 |
| Scapula (AT-6672) | 20 | 150415_SN7001204_0369 | 2x 76 | 1 | 81.9 | 22.8 |
| | | | | 2 | 82.5 | 49.1 |
| | | 150420_SN7001204_0373 | 2x 76 | 1 | 79.6 | 33.7 |
| | | | | 2 | 80.0 | 41.6 |

Provided are the percentage of PhiX spiked into each lane, the number of sequencing cycles performed and the success of index recognition.

doi:10.1038/nature17172

# Memory retrieval by activating engram cells in mouse models of early Alzheimer's disease

Dheeraj S. Roy[1], Autumn Arons[1,2], Teryn I. Mitchell[1], Michele Pignatelli[1], Tomás J. Ryan[1,2] & Susumu Tonegawa[1,2]

**Alzheimer's disease (AD) is a neurodegenerative disorder characterized by progressive memory decline and subsequent loss of broader cognitive functions[1]. Memory decline in the early stages of AD is mostly limited to episodic memory, for which the hippocampus has a crucial role[2]. However, it has been uncertain whether the observed amnesia in the early stages of AD is due to disrupted encoding and consolidation of episodic information, or an impairment in the retrieval of stored memory information. Here we show that in transgenic mouse models of early AD, direct optogenetic activation of hippocampal memory engram cells results in memory retrieval despite the fact that these mice are amnesic in long-term memory tests when natural recall cues are used, revealing a retrieval, rather than a storage impairment. Before amyloid plaque deposition, the amnesia in these mice is age-dependent[3–5], which correlates with a progressive reduction in spine density of hippocampal dentate gyrus engram cells. We show that optogenetic induction of long-term potentiation at perforant path synapses of dentate gyrus engram cells restores both spine density and long-term memory. We also demonstrate that an ablation of dentate gyrus engram cells containing restored spine density prevents the rescue of long-term memory. Thus, selective rescue of spine density in engram cells may lead to an effective strategy for treating memory loss in the early stages of AD.**

AD is the most common cause of brain degeneration, and typically begins with impairments in cognitive functions[1]. Most research has focused on understanding the relationship between memory impairments and the formation of two pathological hallmarks seen in the late stages of AD: extracellular amyloid plaques and intracellular aggregates of tau protein[1,2]. The early phases of AD have received relatively less attention, although synaptic phenotypes have been identified as major correlates of cognitive impairments in both human patients and mouse models[3,6]. Several studies have suggested that the episodic memory deficit of AD patients is due to ineffective encoding of new information[7–9]. However, since the cognitive measures used in these studies rely on memory retrieval, it is not possible to discriminate rigorously between impairments in information storage and disrupted retrieval of stored information. This issue has an important clinical implication: if the amnesia is due to retrieval impairments, memory could be restored by technologies involving targeted brain stimulation.

A mouse model of AD (hereafter referred to as 'AD mice')[10] overexpresses the delta exon 9 variant of presenilin 1 (PS1; also known as PSEN1), in combination with the Swedish mutation of β-amyloid precursor (APP). Consistent with previous reports[3–5], 9-month-old AD mice showed severe plaque deposition across multiple brain regions (Fig. 1a), specifically in the dentate gyrus (DG) (Fig. 1b) and medial entorhinal cortex (EC) (Fig. 1c); in contrast, 7-month-old AD mice lacked amyloid plaques (Fig. 1d and Extended Data Fig. 1a–d). Focusing on these two age groups of AD mice, we quantified short-term (1 h; STM) and long-term (24 h; LTM) memory formation using contextual fear conditioning (CFC) (Fig. 1e). Nine-month-old AD mice were impaired in both STM and LTM, which suggested a deficit in memory encoding (Fig. 1k–o). By contrast, 7-month-old AD mice showed normal levels of training-induced freezing (Fig. 1f) and normal STM (Fig. 1g), but were impaired in LTM (Fig. 1h). Neither control nor 7-month-old AD mice displayed freezing behaviour in a neutral context (Fig. 1i). In the DG of 7-month-old AD mice, the levels of cells that were immediate early gene c-Fos-positive after CFC training were normal, but were lower compared with control mice after LTM tests (Fig. 1j). Motor behaviours and the density of DG granule cells were normal in these mice (Extended Data Fig. 1e–k). Thus, these behavioural- and cellular-level observations confirmed that 7-month-old AD mice serve as a mouse model of early AD regarding memory impairments.

Recently, molecular, genetic and optogenetic methods to identify neurons that hold traces, or engrams, of specific memories have been established[11,12]. Using this technology, several groups have demonstrated that DG neurons activated during CFC learning are both sufficient[11–14] and necessary[15] for subsequent memory retrieval. In addition, our recent study found that engram cells under protein-synthesis-inhibitor-induced amnesia were capable of driving acute memory recall if they were directly activated optogenetically[14]. Here, we applied this memory engram cell identification and manipulation technology to 7-month-old AD mice to determine whether memories could be retrieved in the early stages of the disease. Because it is known that the EC–hippocampus (HPC) network is among the earliest to show altered synaptic/dendritic properties and these alterations have been suggested to underlie the memory deficits in early AD[16,17], we focused on labelling the DG component of CFC memory engram cells of 7-month-old AD mice using a double adeno-associated virus (AAV) system (Fig. 1p, q and Methods). Although on a doxycycline (DOX) diet DG neurons completely lacked channelrhodopsin 2 (ChR2)–enhanced yellow fluorescent protein (eYFP) labelling, 1 day off DOX was sufficient to permit robust ChR2–eYFP expression in control mice (Fig. 1r, s and Extended Data Fig. 2a–c), as well as in 7-month-old AD mice (Fig. 1t, u).

As expected, these engram-labelled early AD mice were amnesic a day after CFC training (Fig. 1v). But, remarkably, these mice froze on the next day in a distinct context (context B) as robustly as equivalently treated control mice in response to blue light stimulation of the engram cells (Fig. 1w). This light-specific freezing was not observed using on-DOX mice (Extended Data Fig. 2d–f). A natural recall test conducted on the third day in the conditioning context (context A) revealed that the observed optogenetic engram reactivation did not restore memory recall by natural cues in early AD mice (Fig. 1x). This was the case even after multiple rounds of light activation of the engram cells (Extended Data Fig. 3). We replicated the successful optogenetic rescue of memory recall in two other models of early AD: a triple transgenic line obtained by mating *c-Fos-tTA* mice with double-transgenic APP/PS1 mice (Extended Data Fig. 4a–g) and a widely used triple-transgenic AD model[18] (PS1/APP/tau (also known as MAPT); Extended Data Fig. 4h–m). These data show that DG engram cells in 7-month-old mouse models of early AD are sufficient to induce memory recall upon optogenetic reactivation, which indicates a deficit of memory retrievability during early AD-related memory loss.

[1]RIKEN-MIT Center for Neural Circuit Genetics at the Picower Institute for Learning and Memory, Department of Biology and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

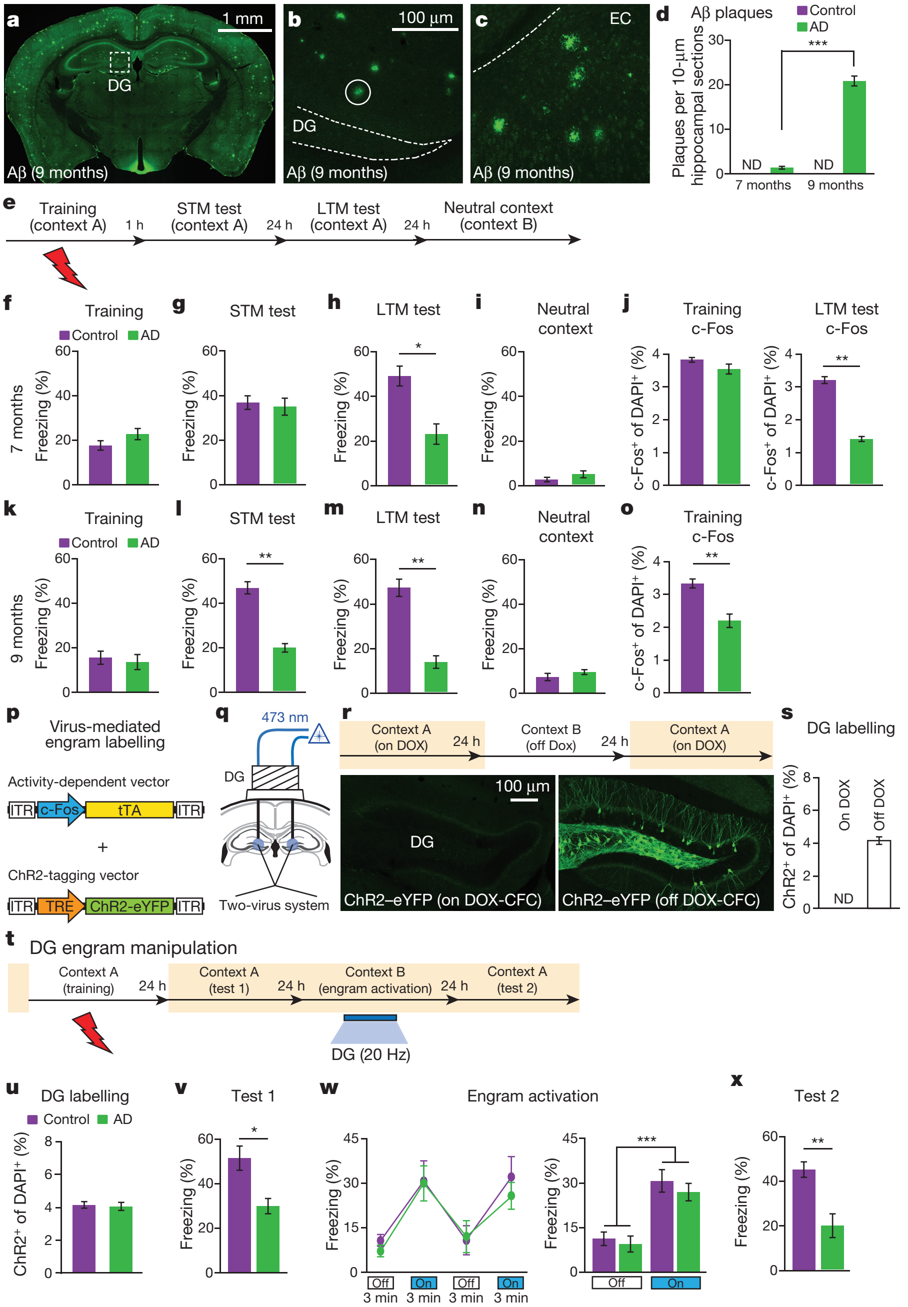


**Figure 1 | Optogenetic activation of memory engrams restores fear memory in early AD mice. a**–**c**, Amyloid-β (Aβ) plaques in 9-month-old AD mice (**a**), in the DG (**b**), and in the EC (**c**). **d**, Plaque counts in HPC sections ($n = 4$ mice per group). ND, not detected. **e**, CFC behavioural schedule ($n = 10$ mice per group). **f**–**i**, Freezing levels of 7-month-old AD groups during training (**f**), STM test (**g**), LTM test (**h**) or exposure to neutral context (**i**). **j**, c-Fos$^+$ cell counts in the DG of 7-month-old mice after CFC training or LTM test, represented in **f**, **h** ($n = 4$ mice per group). DAPI, 4′,6-diamidino-2-phenylindole. **k**–**n**, Freezing levels of 9-month-old AD mice during training (**k**), STM test (**l**), LTM test (**m**) or exposure to neutral context (**n**). **o**, c-Fos$^+$ cell counts in the DG of 9-month-old mice ($n = 3$ mice per group) after CFC training represented in **k**. **p**, Virus-mediated engram labelling strategy using a cocktail of AAV$_9$-c-Fos-tTA and AAV$_9$-TRE-ChR2-eYFP. **q**, AD mice were injected with the two viruses bilaterally and implanted with an optic fibre bilaterally into the DG. **r**, Behavioural schedule and DG engram cell labelling (see Methods). **s**, ChR2–eYFP$^+$ cell counts from DG sections shown in **r** ($n = 3$ mice per group). **t**, Behavioural schedule for optogenetic activation of DG engram cells. **u**, ChR2–eYFP$^+$ cell counts from 7-month-old mice ($n = 5$ mice per group). **v**, Memory recall in context A 1 day after training (test 1, $n = 9$ mice per group). **w**, Freezing by blue light stimulation (left). Average freezing for two light-off and light-on epochs (right). **x**, Memory recall in context A 3 days after training (test 2). Statistical comparisons are performed using unpaired $t$-tests; $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Data are presented as mean ± standard error of the mean (s.e.m.).

Reduced dendritic spines have been implicated in memory impairments of AD[3]. In addition, our recent study of protein-synthesis-inhibitor-induced amnesia found reduced engram-cell-specific dendritic spine density[14]. We detected an age-dependent (Extended Data Fig. 5a) decrease in dendritic spine density of DG engram cells in early AD mice (Fig. 2a–c), showing that the long-term memory impairments of early AD correlate with dendritic spine deficits of DG engram cells (Extended Data Fig. 5b). The inability to generate newborn neurons in the DG could play a part in the development of AD-specific cognitive deficits[19]. However, early AD mice showed similar levels of neurogenesis in the DG compared with control mice, which were quantified using doublecortin (DCX) staining (Extended Data Fig. 1l–q). We recently proposed that the persistent cellular connectivity between multiple engram cell ensembles is a fundamental mechanism of memory information

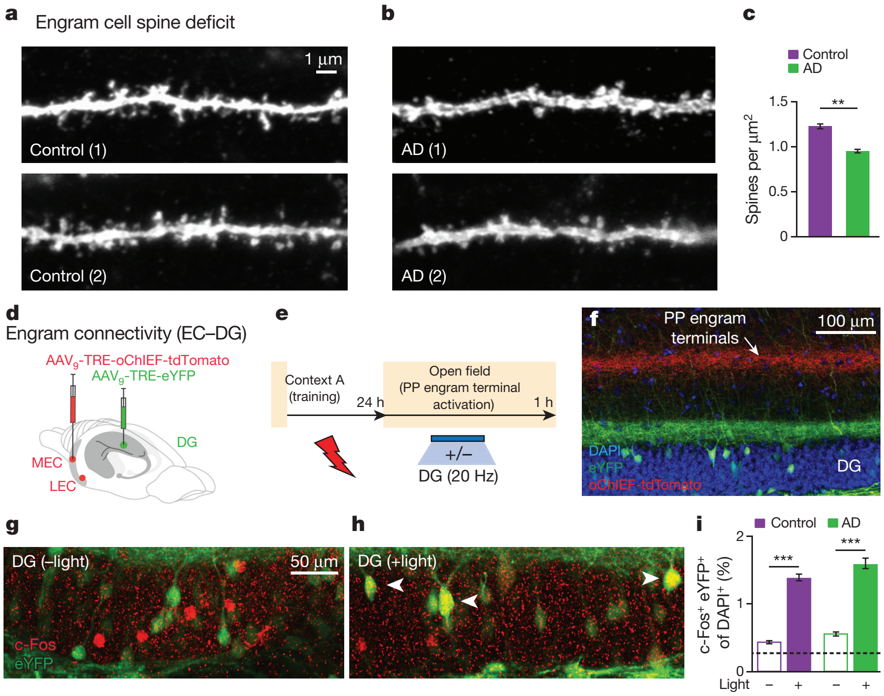


**Figure 2 | Neural correlates of amnesia in early AD mice. a**, **b**, Images showing dendritic spines from DG engram cells of control (**a**) and AD (**b**) groups. **c**, Average spine density showing a decrease in AD mice ($n = 7{,}032$ spines) compared with controls ($n = 9{,}437$ spines, $n = 4$ mice per group). **d**, For engram connectivity, MEC/LEC and DG cells were injected with virus cocktails. **e**, Engram connectivity behavioural schedule. Mice ($n = 4$ per group) were either given a natural exploration session (−) or a PP engram terminal stimulation session (+) in an open field. **f**, Image showing simultaneous labelling of engram terminals (red) and engram cells (green). Green terminals reflect mossy cell axons. **g**, **h**, Images showing c-Fos$^+$/eYFP$^+$ overlap in the DG. **i**, c-Fos$^+$/eYFP$^+$ counts from control and AD mice. Chance overlap (0.24) was calculated (see Methods) and indicated by the dashed line. Statistical comparisons are performed using unpaired $t$-tests; $^{**}P < 0.01$, $^{***}P < 0.001$. Data are presented as mean ± s.e.m.

retention[14]. We labelled putative CFC memory engram cells in both medial EC (MEC) and lateral EC (LEC) with oChIEF[20] (a variant of ChR2) and simultaneously labelled CFC memory engram cells in the DG with eYFP (Fig. 2d). With this procedure, perforant path (PP) terminals are also labelled with oChIEF (Fig. 2e, f). One day after footshocks, we optogenetically activated these terminals and quantified the overlap between putative DG engram cells (that is, eYFP$^+$, green) and DG cells in which the endogenous c-Fos (red) had been activated by the optogenetic activation of oChIEF$^+$ PP terminals. Both control and early AD mice showed above-chance and indistinguishable levels of c-Fos$^+$/eYFP$^+$ overlap, indicating that the preferential functional connectivity between engram cells is maintained in the early AD mice (Fig. 2g–i).

We then hypothesized that the reversal of dendritic spine deficits in DG engram cells of early AD mice may rescue long-term memory. To investigate this possibility, we took advantage of previous findings that spine formation can be induced rapidly by long-term potentiation (LTP)[21,22] and that LTP can be induced *in vivo* using light activation of oChIEF[23]. We validated learning-dependent labelling, with oChIEF, of neurons in the MEC (Fig. 3a–c and Extended Data Fig. 6a–c) and LEC (Fig. 3d) as well as PP terminals in the DG (Fig. 3e, f). *In vivo* extracellular recording upon light stimulation of oChIEF$^+$ EC axonal terminals in the DG showed a reliable spiking response of DG cells in anaesthetized control mice (Fig. 3g). Furthermore, in HPC slices from control mice we successfully induced LTP in DG cells using a previously established optical LTP protocol[23] (Fig. 3h–j). These biocytin-filled DG cells revealed an increase in spine density after *in vitro* optical LTP (Extended Data Fig. 6d).

In early AD mice, *in vivo* application of the engram-specific optical LTP protocol restored spine density of DG engram cells to control levels (AD + 100 Hz group; Fig. 3k, l). Furthermore, this spine restoration in early AD mice correlated with amelioration of long-term memory impairments observed during recall by natural cues (Fig. 3m), an effect that persisted for at least 6 days after training (AD rescue + diphtheria toxin receptor (DTR) + saline group; Fig. 3p). The LTP-induced spine restoration and behavioural deficit rescue were protein-synthesis dependent (Extended Data Fig. 7). The rescued memory was context-specific (Extended Data Fig. 8a). In addition, long-term memory recall of age-matched control mice was unaffected by this optical LTP protocol (Extended Data Fig. 8b). By contrast, applying the optical LTP protocol to a large portion of excitatory PP terminals in the DG (that is, with no restriction to the PP terminals derived from EC engram cells) did not result in long-term memory rescue in early AD mice (Extended Data Fig. 9). To confirm the correlation between restoration of spine density of DG engram cells and amelioration of long-term memory impairments, which were both induced by the optical LTP protocol, we compared the overlap of natural-recall-cue-induced c-Fos$^+$ cells and CFC-training-labelled DG engram cells after an application of the engram-specific LTP protocol to early AD mice (Fig. 3n). Early AD mice that did not receive the optical LTP protocol showed low levels of c-Fos$^+$/eYFP$^+$ overlap compared with control mice upon natural recall cue delivery. By contrast, early AD mice that went through the optical LTP protocol showed c-Fos$^+$/eYFP$^+$ overlap similar to that of control mice (Fig. 3n). Thus, these data suggest that spine density restoration in DG engram cells contributes to the rescue of long-term memory in early AD mice.

Because of the highly redundant connectivity between the EC and DG[24], it is possible that the extensive optical LTP protocol also augmented spine density in some non-engram DG cells. To establish a link between the spine rescue in DG engram cells and the behavioural rescue of early AD mice, we developed an engram-specific ablation[25] virus. We confirmed that this DTR-mediated method efficiently ablated DG engram cells after diphtheria toxin (DT) administration (Fig. 3o), while leaving the nearby DG mossy cells intact (Extended Data Fig. 10). By simultaneously labelling axonal terminals of PP with oChIEF and DG engram cells with DTR, we examined the effect of DG engram cell ablation after optical LTP-induced behavioural rescue (Fig. 3p). Within-animal comparisons (test 1 versus test 2) showed a decrease in freezing behaviour of LTP-rescued AD mice in which DG engram cells were ablated. These data strengthen the link between DG engram cells with restored spine density and long-term behavioural rescue in early AD mice.

To examine whether the optical LTP-induced behavioural rescue could be applied to DG engram cells from other learning experiences, we labelled memory engrams for inhibitory avoidance or novel object location in early AD mice (Fig. 4a). Early AD mice showed memory impairments in inhibitory avoidance memory and novel object location spatial memory (Fig. 4b, c). Optical LTP-induced spine rescue at the PP–DG engram synapses was sufficient to reverse long-term memory impairments of early AD mice in both behavioural paradigms, thus demonstrating the versatility of our engram-based intervention.

Previous studies that examined the early stages of AD found correlations between memory impairments and synaptic pathology at the

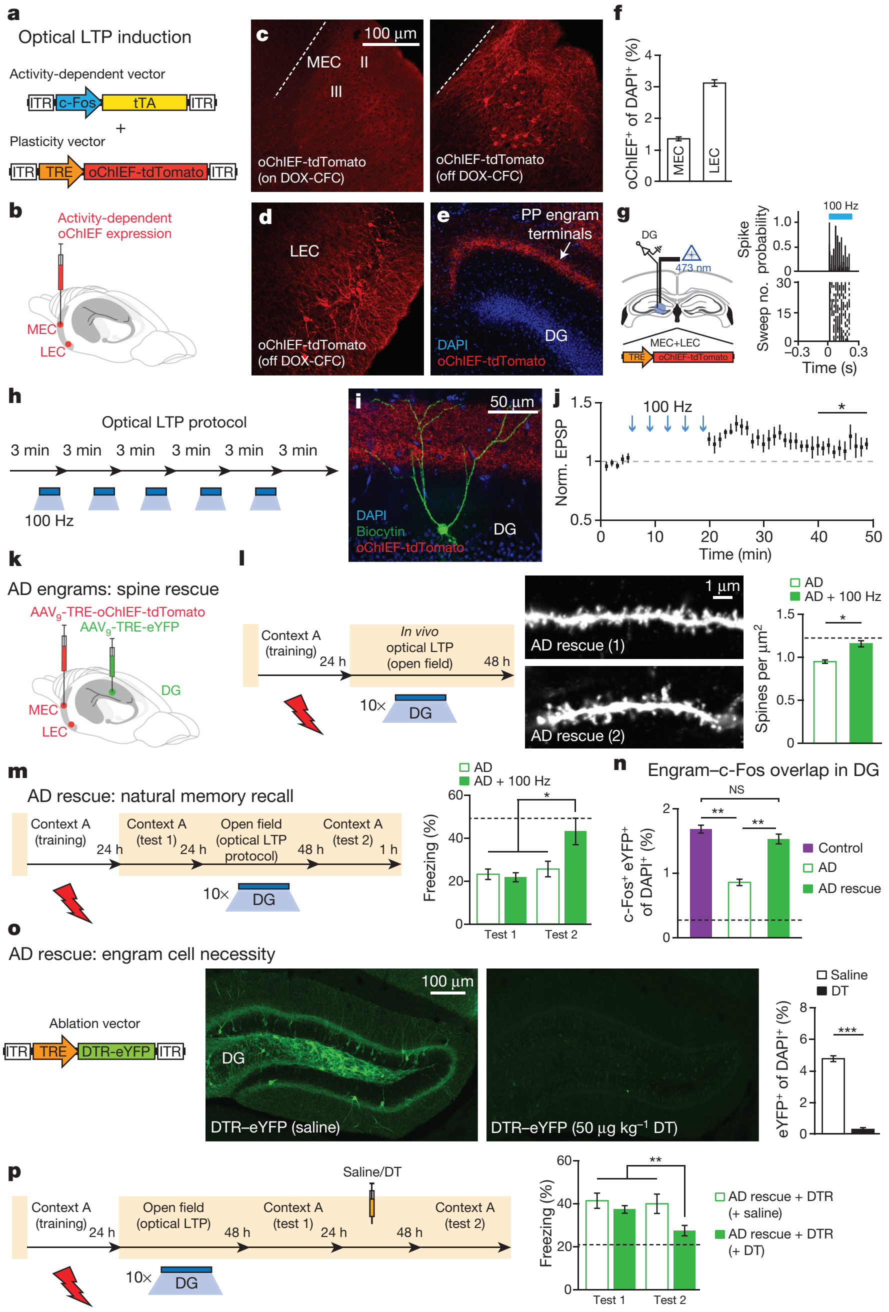


**Figure 3 | Reversal of engram-specific spine deficits rescues memory in early AD mice.** **a**, Engram-specific optical LTP using two viruses. **b**, Virus cocktail injected into MEC/LEC. **c**–**e**, Images showing oChIEF labelling 24 h after CFC: in MEC on DOX (left; **c**) and off DOX (right; **c**); in LEC off DOX (**d**); in DG off DOX (sagittal; **e**). Scale bar shown in **c** applies to **d** and **e**. **f**, oChIEF$^+$ cell counts ($n = 3$ mice per group). **g**, *In vivo* spiking of DG neurons in response to 100 Hz light applied to PP terminals. **h**, Optical LTP protocol[23]. **i**, **j**, *In vitro* responses of DG cells after optical LTP. Image showing biocytin-filled DG cell receiving oChIEF$^+$ PP terminals (coronal; **i**). Normalized (Norm.) excitatory post-synaptic potentials (EPSPs) showing a 10% increase in amplitude ($n = 6$ cells; **j** and Methods). **k**, For *in vivo* optical LTP at EC–DG synapses, MEC/LEC and DG cells were injected with virus cocktails. **l**, Protocol for *in vivo* spine restoration of DG engram cells in AD mice (left). Images showing dendritic spines of DG engram cells after LTP (middle). A two-way analysis of variance (ANOVA) followed by Bonferroni post-hoc tests revealed a spine density restoration in AD + 100 Hz mice ($F_{1,211} = 7.21$, $P < 0.01$, 13,025 spines, $n = 4$ mice per group; right). Dashed line represents control mice spine density (1.21). **m**, Behavioural schedule for memory rescue in AD mice (left). A two-way ANOVA with repeated measures followed by Bonferroni post-hoc tests revealed restored freezing in AD + 100 Hz mice ($F_{1,36} = 4.95$, $P < 0.05$, $n = 10$ mice per group; right). Dashed line represents control mice freezing (48.53). **n**, After rescue, mice were perfused for c-Fos$^+$/eYFP$^+$ overlap cell counts. Chance was estimated at 0.22. NS, not significant. **o**, Construct for ablation of engram cells using DTR (left). Images showing DG engram cells after saline/DT administration (middle). DTR–eYFP cell counts ($n = 5$ mice per group; right). **p**, Behavioural schedule testing the necessity of engram cells after spine restoration (left). Memory recall showed less freezing of AD mice treated with DT (AD rescue + DTR + DT) compared with saline-treated mice ($n = 9$ mice per group; right). Dashed line represents freezing of non-stimulated early AD mice (20.48). Unless specified, statistical comparisons are performed using unpaired $t$-tests; $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. Data are presented as mean ± s.e.m.

EC PP input into the DG[3,4,6]. It has been proposed that these early cognitive deficits are a failure of memory encoding on the basis of behavioural observations in human patients[8,9]. However, we have shown that optogenetic activation of HPC cells active during learning elicits memory recall in mouse models of early AD. To our knowledge, this is the first rigorous demonstration that memory failure in early AD models reflects an impairment in the retrieval of information. Further support for a memory retrieval impairment in early AD comes from the fact that impairments are in LTM (at least 1 day long), but not in STM (~1 h after training), which is consistent with a retrieval deficit. The retrieval deficit in early AD models is similar to memory deficits observed in amnesia induced by impairing memory consolidation via protein synthesis inhibitors[14]. The underlying mechanism of memory failure in early AD patients may not necessarily parallel the molecular and circuit impairments observed in mouse models of early AD. For instance, some early AD patients can exhibit amyloid plaque deposition years before the onset of cognitive decline[9]. However, converging data on the underlying mechanism for genetically and pharmacologically induced amnesia in animal models increase the possibility that similar memory-retrieval-based failures may also operate in an early stage of AD patients. While we have shown that amnesia in early AD mice is a deficit of memory retrieval, it remains possible that the long-term maintenance of memory storage may also gradually become compromised as the disease proceeds from the early stage to more advanced stages, and eventually lost with neuronal degeneration. Further research will investigate these possibilities.

Our conclusions apply to episodic memory, which involves processing by HPC and other medial temporal lobe structures. In the literature[9], it is widely recognized that early AD patients exhibit non-episodic memory deficits as well, which would involve brain structures other than the medial temporal lobe. Additional work is required to examine the mechanisms underlying cognitive impairments in these other types of memories. Nevertheless, our findings already contribute to a better understanding of memory retrieval deficits in several cases of early AD, and may apply to other pathological conditions, such as Huntington's disease[8], in which patients show difficulty in memory recall.

Consistent with several studies highlighting the importance of dendritic spines[3,6,14,26] in relation to memory processing, we observed an engram-cell-specific decrease in spine density that correlated with

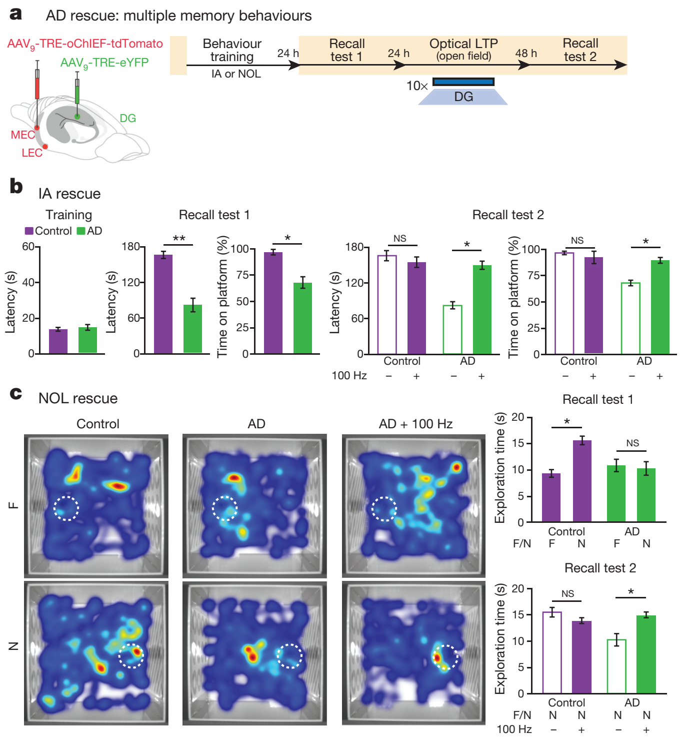


**Figure 4 | Recovery of multiple types of HPC-dependent memories from amnesia in early AD. a**, MEC/LEC and DG cells were injected with virus cocktails (left). Behavioural schedule for engram labelling (right). **b**, Inhibitory avoidance (IA) long-term rescue ($n = 10$ mice per group). Recall test 1 showed decreased latency and time on platform for AD mice. A two-way ANOVA with repeated measures followed by Bonferroni post-hoc tests revealed a recovery of IA memory in early AD mice (latency: $F_{1,27} = 25.22$, $P < 0.001$; time on platform: $F_{1,27} = 6.46$, $P < 0.05$; recall test 2). **c**, Novel object location (NOL) long-term rescue ($n = 15$ mice per group). Average heat maps showing exploration time for familiar (F) or novel (N) locations (left or right, respectively). White circles represent object location. Recall test 1 showed comparable exploration of familiar locations by control and AD mice; however, AD mice showed decreased exploration of novel locations. A two-way ANOVA with repeated measures followed by Bonferroni post-hoc tests revealed a recovery of NOL memory in early AD mice ($F_{1,56} = 5.87$, $P < 0.05$; recall test 2). Unless specified, statistical comparisons are performed using unpaired $t$-tests; $*P < 0.05$, $**P < 0.01$. Data are presented as mean ± s.e.m. NS, not significant.

memory deficits in early AD. Natural rescue of memory recall in early AD mice required the DG engram cells in which synaptic density deficits have been restored by *in vivo* optical LTP protocols applied to the EC cells activated during learning. By contrast, the application of optical LTP protocols to a much wider array of excitatory EC cells projecting to the DG, which may be analogous to deep brain stimulation, did not rescue memory in AD mice. A potential explanation for this observation is that DG granule cells may contribute to a variety of memories through their partially overlapping engram cell ensembles in a competitive manner, and that activation of a large number of these ensembles simultaneously may interfere with a selective activation of an individual ensemble. Thus, activation of a more targeted engram cell ensemble may be a key requirement for effective retrieval of the specific memory, which is difficult to achieve with the current deep brain stimulation strategy.

Genetic manipulations of specific neuronal populations can have profound effects on cognitive impairments of AD[27]. We propose that strategies applied to engram circuits can support long-lasting improvements in cognitive functions, which may provide insights and therapeutic value for future approaches that rescue memory in AD patients.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 6 August 2015; accepted 28 January 2016.**
**Published online 16 March 2016.**

1. Selkoe, D. J. Alzheimer's disease: genes, proteins, and therapy. *Physiol. Rev.* **81,** 741–766 (2001).
2. Selkoe, D. J. Alzheimer's disease is a synaptic failure. *Science* **298,** 789–791 (2002).
3. Jacobsen, J. S. *et al.* Early-onset behavioral and synaptic deficits in a mouse model of Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **103,** 5161–5166 (2006).
4. Hsia, A. Y. *et al.* Plaque-independent disruption of neural circuits in Alzheimer's disease mouse models. *Proc. Natl Acad. Sci. USA* **96,** 3228–3233 (1999).
5. Mucke, L. *et al.* High-level neuronal expression of A$\beta_{1-42}$ in wild-type human amyloid protein precursor transgenic mice: synaptotoxicity without plaque formation. *J. Neurosci.* **20,** 4050–4058 (2000).
6. Terry, R. D. *et al.* Physical basis of cognitive alterations in Alzheimer's disease: synapse loss is the major correlate of cognitive impairment. *Ann. Neurol.* **30,** 572–580 (1991).
7. Granholm, E. & Butters, N. Associative encoding and retrieval in Alzheimer's and Huntington's disease. *Brain Cogn.* **7,** 335–347 (1988).
8. Hodges, J. R., Salmon, D. P. & Butters, N. Differential impairment of semantic and episodic memory in Alzheimer's and Huntington's diseases: a controlled prospective study. *J. Neurol. Neurosurg. Psychiatry* **53,** 1089–1095 (1990).
9. Weintraub, S., Wicklund, A. H. & Salmon, D. P. The neuropsychological profile of Alzheimer disease. *Cold Spring Harb. Perspect. Med.* **2,** a006171 (2012).
10. Jankowsky, J. L. *et al.* Mutant presenilins specifically elevate the levels of the 42 residue β-amyloid peptide *in vivo*: evidence for augmentation of a 42-specific γ secretase. *Hum. Mol. Genet.* **13,** 159–170 (2004).
11. Liu, X. *et al.* Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature* **484,** 381–385 (2012).
12. Ramirez, S. *et al.* Creating a false memory in the hippocampus. *Science* **341,** 387–391 (2013).
13. Redondo, R. L. *et al.* Bidirectional switch of the valence associated with a hippocampal contextual memory engram. *Nature* **513,** 426–430 (2014).
14. Ryan, T. J., Roy, D. S., Pignatelli, M., Arons, A. & Tonegawa, S. Engram cells retain memory under retrograde amnesia. *Science* **348,** 1007–1013 (2015).
15. Denny, C. A. *et al.* Hippocampal memory traces are differentially modulated by experience, time, and adult neurogenesis. *Neuron* **83,** 189–201 (2014).
16. Harris, J. A. *et al.* Transsynaptic progression of amyloid-β-induced neuronal dysfunction within the entorhinal-hippocampal network. *Neuron* **68,** 428–441 (2010).
17. Hyman, B. T., Van Hoesen, G. W., Kromer, L. J. & Damasio, A. R. Perforant pathway changes and the memory impairment of Alzheimer's disease. *Ann. Neurol.* **20,** 472–481 (1986).
18. Oddo, S. *et al.* Triple-transgenic model of Alzheimer's disease with plaques and tangles: intracellular Aβ and synaptic dysfunction. *Neuron* **39,** 409–421 (2003).
19. Rodríguez, J. J. *et al.* Impaired adult neurogenesis in the dentate gyrus of a triple transgenic mouse model of Alzheimer's disease. *PLoS One* **3,** e2935 (2008).
20. Lin, J. Y., Lin, M. Z., Steinbach, P. & Tsien, R. Y. Characterization of engineered channelrhodopsin variants with improved properties and kinetics. *Biophys. J.* **96,** 1803–1814 (2009).
21. Maletic-Savatic, M., Malinow, R. & Svoboda, K. Rapid dendritic morphogenesis in CA1 hippocampal dendrites induced by synaptic activity. *Science* **283,** 1923–1927 (1999).
22. Engert, F. & Bonhoeffer, T. Dendritic spine changes associated with hippocampal long-term synaptic plasticity. *Nature* **399,** 66–70 (1999).
23. Nabavi, S. *et al.* Engineering a memory with LTD and LTP. *Nature* **511,** 348–352 (2014).
24. Tamamaki, N. & Nojyo, Y. Projection of the entorhinal layer II neurons in the rat as revealed by intracellular pressure-injection of neurobiotin. *Hippocampus* **3,** 471–480 (1993).
25. Zhan, C. *et al.* Acute and long-term suppression of feeding behavior by POMC neurons in the brainstem and hypothalamus, respectively. *J. Neurosci.* **33,** 3624–3632 (2013).
26. Tonegawa, S., Liu, X., Ramirez, S. & Redondo, R. Memory engram cells have come of age. *Neuron* **87,** 918–931 (2015).
27. Cissé, M. *et al.* Reversing EphB2 depletion rescues cognitive functions in Alzheimer model. *Nature* **469,** 47–52 (2011).

**Acknowledgements** We thank X. Liu for the c-Fos-tTA construct; S. Huang, T. Okuyama and T. Kitamura for help with experiments; W. Yu, S. LeBlanc and X. Zhou for technical assistance; L. Brenner for proofreading; and all members of the Tonegawa laboratory for their support. We thank M. Luo for sharing the DTR coding sequence. This work was supported by the RIKEN Brain Science Institute, the Howard Hughes Medical Institute, and the JPB Foundation (to S.T.).

**Author Contributions** D.S.R. and S.T. contributed to the study design. D.S.R., A.A., T.I.M., M.P. and T.J.R. contributed to the data collection and interpretation. D.S.R. cloned all constructs. D.S.R. and A.A. conducted the surgeries, behaviour experiments and histological analyses. D.S.R. and S.T. wrote the paper. All authors discussed and commented on the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.T. (tonegawa@mit.edu).

## METHODS

**Subjects.** The APP/PS1 (ref. 10) double-transgenic AD mice, originally described as Line 85, were obtained from Jackson Laboratory (stock number 004462). Under the control of mouse prion promoter elements, these mice express a chimaeric mouse/human *APP* transgene containing Swedish mutations (K595N/M596L) as well as a mutant human *PS1* transgene (delta exon 9 variant). To label memory engram cells in APP/PS1 mice, we generated a triple-transgenic mouse line by mating *c-Fos-tTA*[11,28] transgenic mice with APP/PS1 double-transgenic mice. The PS1/APP/tau[18] triple-transgenic AD mice were obtained from Jackson Laboratory (stock number 004807). These 3×Tg-AD mice express a mutant human *PS1* transgene (M146V), a human *APP* transgene containing Swedish mutations (KM670/671NL) and a human *MAPT* transgene harbouring the P301L mutation. All mouse lines were maintained as hemizygotes. Mice had access to food and water *ad libitum* and were socially housed in numbers of two to five littermates until surgery. After surgery, mice were singly housed. For behavioural experiments, all mice were male and 7–9 months old. For optogenetic experiments, mice had been raised on food containing 40 mg kg$^{-1}$ DOX for at least 1 week before surgery, and remained on DOX for the remainder of the experiments except for the target engram labelling days. For *in vitro* electrophysiology experiments, mice were 24–28 days old at the time of surgery. All experiments were conducted in accordance with US National Institutes of Health (NIH) guidelines and the Massachusetts Institute of Technology Department of Comparative Medicine and Committee of Animal Care. No statistical methods were used to predetermine sample size.

**Viral constructs.** Our previously established method[11] for labelling memory engram cells combined c-Fos-tTA transgenic mice with a DOX-sensitive adeno-associated virus (AAV). However, in this study, we modified the method using a double-virus system to label memory engram cells in the early AD mice, which already carry two transgenes. The pAAV-c-Fos-tTA plasmid was constructed by cloning a 1 kb fragment from the *c-Fos* gene (550 bp upstream of *c-Fos* exon I to 35 bp into exon II) into an AAV backbone using the KpnI restriction site at the 5′ terminus and the SpeI restriction site at the 3′ terminus. The AAV backbone contained the tTA-Advanced[29] sequence at the SpeI restriction site. The pAAV-TRE-ChR2-eYFP and pAAV-TRE-eYFP constructs were previously described[11,12]. The pAAV-TRE-oChIEF-tdTomato[20] plasmid was constructed by replacing the ChR2-eYFP fragment from the pAAV-TRE-ChR2-eYFP plasmid using NheI and MfeI restriction sites. The pAAV-CaMKII-oChIEF-tdTomato plasmid was constructed by replacing the TRE fragment from the pAAV-TRE-oChIEF-tdTomato plasmid using BamHI and EcoRI restriction sites. The pAAV-TRE-DTR-eYFP[25] plasmid was constructed by replacing the ChR2 fragment from the pAAV-TRE-ChR2-eYFP plasmid using EcoRI and AgeI restriction sites. AAV vectors were serotyped with AAV$_9$ coat proteins and packaged at the University of Massachusetts Medical School Gene Therapy Center and Vector Core. Viral titres were $1.5 \times 10^{13}$ genome copy (GC) ml$^{-1}$ for AAV$_9$-c-Fos-tTA, AAV$_9$-TRE-ChR2-eYFP and AAV$_9$-TRE-eYFP, $1 \times 10^{13}$ GC ml$^{-1}$ for AAV$_9$-TRE-oChIEF-tdTomato, $4 \times 10^{13}$ GC ml$^{-1}$ for AAV$_9$-CaMKII-oChIEF-tdTomato and $2 \times 10^{13}$ GC ml$^{-1}$ for AAV$_9$-TRE-DTR-eYFP.

**Surgery and optic fibre implants.** Mice were anaesthetized with isoflurane or 500 mg kg$^{-1}$ avertin for stereotaxic injections[14]. Injections were targeted bilaterally to the DG (−2.0 mm anteroposterior (AP), ±1.3 mm mediolateral (ML), −1.9 mm dorsoventral (DV)), MEC (−4.7 mm AP, ±3.35 mm ML, −3.3 mm DV) and LEC (−3.4 mm AP, ±4.3 mm ML, −4.0 mm DV). Injection volumes were 300 nl for DG and 400 nl for MEC and LEC. Viruses were injected at 70 nl min$^{-1}$ using a glass micropipette attached to a 10 ml Hamilton microsyringe. The needle was lowered to the target site and remained for 5 min before beginning the injection. After the injection, the needle stayed for 10 min before it was withdrawn. A custom DG implant containing two optic fibres (200 mm core diameter; Doric Lenses) was lowered above the injection site (−2.0 mm AP, ±1.3 mm ML, −1.7 mm DV). The implant was secured to the skull with two jewellery screws, adhesive cement (C&B Metabond) and dental cement. An opaque cap derived from the top part of an Eppendorf tube protected the implant. Mice were given 1.5 mg kg$^{-1}$ metacam as analgesic and allowed to recover for 2 weeks before behavioural experiments. All injection sites were verified histologically. As criteria, we only included mice with virus expression limited to the targeted regions.

**Systemic injection of kainic acid.** For seizure experiments[11], mice were taken off DOX for 1 day and injected intraperitoneally with 15 mg kg$^{-1}$ kainic acid (KA). Mice were returned to DOX food 6 h after KA treatment and perfused the next day for immunohistochemistry procedures.

**Immunohistochemistry.** Mice were dispatched using 750–1,000 mg kg$^{-1}$ avertin and perfused transcardially with PBS, followed by 4% paraformaldehyde (PFA). Brains were extracted and incubated in 4% PFA at room temperature overnight. Brains were transferred to PBS and 50-μm coronal slices were prepared using a vibratome. For immunostaining[14], each slice was placed in PBS + 0.2% Triton X-100 (PBS-T), with 5% normal goat serum for 1 h and then incubated with primary antibody at 4 °C for 24 h. Slices then underwent three wash steps for 10 min each in PBS-T, followed by 1 h incubation with secondary antibody. After three more wash steps of 10 min each in PBS-T, slices were mounted on microscope slides. All analyses were performed blind to the experimental conditions. Antibodies used for staining were as follows: to stain for ChR2–eYFP, DTR–eYFP or eYFP alone, slices were incubated with primary chicken anti-GFP (1:1,000, Life Technologies) and visualized using anti-chicken Alexa-488 (1:200). For plaques, slices were stained using primary mouse anti-β-amyloid (1:1,000; Sigma-Aldrich) and secondary anti-mouse Alexa-488 (1:500). c-Fos was stained with rabbit anti-c-Fos (1:500, Calbiochem) and anti-rabbit Alexa-568 (1:300). Adult newborn neurons were stained with guinea pig anti-DCX (1:1,000; Millipore) and anti-guinea-pig Alexa-555 (1:500). Neuronal nuclei were stained with mouse anti-NeuN (1:200; Millipore) and Alexa-488 (1:200). DG mossy cell axons were stained with mouse anti-CR (1:1,000; Swant) and Alexa-555 (1:300).

**Cell counting.** To characterize the expression pattern of ChR2–eYFP, DTR–eYFP, eYFP alone and oChIEF-tdTomato in control and AD mice, the number of eYFP$^{+}$/tdTomato$^{+}$ neurons were counted from 4–5 coronal slices per mouse ($n = 3$–5 mice per group). Coronal slices centred on coordinates covered by optic fibre implants were taken for DG quantification and sagittal slices centred on injection coordinates were taken for MEC and LEC. Fluorescence images were acquired using a Zeiss AxioImager.Z1/ApoTome microscope (×20). Automated cell counting analysis was performed using ImageJ software. The cell body layers of DG granule cells (upper blade), MEC or LEC cells were outlined as a region of interest (ROI) according to the DAPI signal in each slice. The number of eYFP$^{+}$/tdTomato$^{+}$ cells per section was calculated by applying a threshold above background fluorescence. Data were analysed using Microsoft Excel with the Statplus plug-in. A similar approach was applied for quantifying amyloid-β plaques, c-Fos$^{+}$ neurons and adult newborn (DCX$^{+}$) neurons. Total engram cell reactivation was calculated as ((c-Fos$^{+}$ eYFP$^{+}$)/(total DAPI$^{+}$)) × 100. Chance overlap was calculated as ((c-Fos$^{+}$/total DAPI$^{+}$) × (eYFP$^{+}$/total DAPI$^{+}$)) × 100. Percentage of adult newborn neurons expressing neuronal markers was calculated as ((NeuN$^{+}$ DCX$^{+}$)/(total DCX$^{+}$) × 100. DAPI$^{+}$ counts were approximated from five coronal/sagittal slices using ImageJ. All counting experiments were conducted blind to experimental group. Researcher 1 trained the animals, prepared slices and randomized images, while researcher 2 performed semi-automated cell counting. Statistical comparisons were performed using unpaired *t*-tests: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.

**Spine density analysis.** Engram cells were labelled using c-Fos-tTA-driven synthesis of ChR2–eYFP or eYFP alone. The eYFP signal was amplified using immunohistochemistry procedures, after which fluorescence *z*-stacks were taken by confocal microscopy (Zeiss LSM700) using a ×40 objective. Maximum intensity projections were generated using ZEN Black software (Zeiss). Four mice per experimental group were analysed for dendritic spines. For each mouse, 30–40 dendritic fragments of 10-μm length were quantified ($n = 120$–160 fragments per group). To measure spine density of DG engram cells with a focus on entorhinal cortical inputs, distal dendritic fragments in the middle-to-outer molecular layer (ML) were selected. For CA3 and CA1 engram cells, apical and basal dendritic fragments were selected. To compute spine density, the number of spines counted on each fragment was normalized by the cylindrical approximation of the surface of the specific fragment. Experiments were conducted blind to experimental group. Researcher 1 imaged dendritic fragments and randomized images, while researcher 2 performed manual spine counting.

***In vitro* recordings.** After isoflurane anaesthesia, brains were quickly removed and used to prepare sagittal slices (300 μm) in an oxygenated cutting solution at 4 °C with a vibratome[14]. Slices were incubated at room temperature in oxygenated artificial cerebrospinal fluid (ACSF) until the recordings. The cutting solution contained (in mM): 3 KCl, 0.5 CaCl$_2$, 10 MgCl$_2$, 25 NaHCO$_3$, 1.2 NaH$_2$PO$_4$, 10 D-glucose, 230 sucrose, saturated with 95% O$_2$–5% CO$_2$ (pH 7.3, osmolarity of 340 mOsm). The ACSF contained (in mM): 124 NaCl, 3 KCl, 2 CaCl$_2$, 1.3 MgSO$_4$, 25 NaHCO$_3$, 1.2 NaH$_2$PO$_4$, 10 D-glucose, saturated with 95% O$_2$–5% CO$_2$ (pH 7.3, 300 mOsm). Individual slices were transferred to a submerged experimental chamber and perfused with oxygenated ACSF warmed at 35 °C (±0.5 °C) at a rate of 3 ml min$^{-1}$ during recordings. Current or voltage clamp recordings were performed under an IR-DIC microscope (Olympus) with a ×40 water immersion objective (0.8 NA), equipped with four automatic manipulators (Luigs & Neumann) and a CCD camera (Hamamatsu). Borosilicate glass pipettes (Sutter Instruments) were fabricated with resistances of 8–10 MΩ. The intracellular solution (in mM) for current clamp recordings was: 110 K-gluconate, 10 KCl, 10 HEPES, 4 ATP, 0.3 GTP, 10 phosphocreatine, 0.5% biocytin (pH 7.25, 290 mOsm). Recordings used two dual channel amplifiers (Molecular Devices), a 2 kHz filter, 20 kHz digitization and an ADC/DAC data acquisition unit (Instrutech) running on custom software in Igor Pro (Wavemetrics). Data acquisition was suspended whenever the resting membrane potential was depolarized above −50 mV or the access resistance (RA) exceeded 20 MΩ. Optogenetic stimulation was achieved using a 460 nm LED light source (Lumen Dynamics) driven by TTL input with a delay onset of 25 µs (subtracted offline for latency estimation). Light power on the sample was 33 mW mm$^{-2}$. To test oChIEF expression, EC cells were stimulated with a single light pulse of 1 s, repeated 10 times every 5 s. DG granule cells were held at −70 mV. Optical LTP protocol: 5 min baseline (10 blue light pulses of 2 ms each, repeated every 30 s) was acquired before the onset of the LTP protocol (100 blue light pulses of 2 ms each at a frequency of 100 Hz, repeated 5 times every 3 min) and the effect on synaptic amplitude was recorded for 30 min (1 pulse of 2 ms every 30 s). Using the 5 min baseline recording data, EPSPs were normalized (Fig. 3j). Potentiation was observed in 6 out of 30 cells and results were statistically confirmed using a two-tailed paired *t*-test. Experiments were performed in the presence of 10 µM gabazine (Tocris) and 2 µM CGP55845 (Tocris). Recorded cells were recovered for morphological identification using streptavidin CF633 (Biotium).

***In vivo* recordings.** Multi-unit responses to optical stimulation were recorded in the DG of mice injected with a cocktail of AAV$_9$-c-Fos-tTA and AAV$_9$-TRE-oChIEF-tdTomato viruses into MEC/LEC. Mice were anaesthetized (10 ml kg$^{-1}$) using a mixture of ketamine (100 mg ml$^{-1}$)/xylazine (20 mg ml$^{-1}$) and placed in the stereotactic system. Anaesthesia was maintained by booster doses of ketamine (100 mg kg$^{-1}$). An optrode consisting of a tungsten electrode (0.5 MΩ) attached to an optic fibre (200-µm core diameter), with the tip of the electrode extending beyond the tip of the fibre by 300 µm, was used for simultaneous optical stimulation and extracellular recording. The power intensity of light emitted from the optrode was calibrated to about 10 mW, consistent with the power used in behavioural assays. oChIEF$^+$ cells were identified by delivering 20-ms light pulses (1 Hz) to the recording site every 50–100 µm. After light-responsive cells were detected, multi-unit activity in response to trains of light pulses (200 ms) at 100 Hz was recorded. Data acquisition used an Axon CNS Digidata 1440A system. MATLAB analysis was performed, as previously described[12].

**Behaviour assays.** Experiments were conducted during the light cycle (7 a.m. to 7 p.m.). Mice were randomly assigned to experimental groups for specific behavioural assays immediately after surgery. Mice were habituated to investigator handling for 1–2 min on three consecutive days. Handling took place in the holding room where the mice were housed. Before each handling session, mice were transported by wheeled cart to and from the vicinity of the behaviour rooms to habituate them to the journey. For natural memory recall sessions, data were quantified using FreezeFrame software. Optogenetic stimulation interfered with the motion detection, and therefore all light-induced freezing behaviour was manually quantified. All behaviour experiments were analysed blind to experimental group. Unpaired Student's *t*-tests were used for independent group comparisons, with Welch's correction when group variances were significantly different. Given behavioural variability, initial assays were performed using a minimum of 10 mice per group to ensure adequate power for any observed differences. Experiments that resulted in significant behavioural effects were replicated three times in the laboratory. Following behavioural protocols, brain sections were prepared to confirm efficient viral labelling in target areas. Animals lacking adequate labelling were excluded before behaviour quantification.

**Contextual fear conditioning.** Two distinct contexts were employed[14]. Context A was 29 × 25 × 22 cm chambers with grid floors, opaque triangular ceilings, red lighting, and scented with 1% acetic acid. Four mice were run simultaneously in four identical context A chambers. Context B consisted of four 30 × 25 × 33 cm chambers with perspex floors, transparent square ceilings, bright white lighting, and scented with 0.25% benzaldehyde. All mice were conditioned in context A (two 0.60 mA shocks of 2 s duration in 5 min), and tested (3 min) in contexts A and B 1 day later. Experiments showed no generalization in the neutral context B. All experimental groups were counter-balanced for chamber within contexts. Floors of chambers were cleaned with quatricide before and between runs. Mice were transported to and from the experimental room in their home cages using a wheeled cart. The cart and cages remained in an anteroom to the experimental rooms during all behavioural experiments. For engram labelling, mice were kept on regular food without DOX for 24 h before training. When training was complete, mice were switched back to food containing 40 mg kg$^{-1}$ DOX.

**Open field.** Spontaneous motor activity was measured in an open field arena (52 × 26 cm) for 10 min. All mice were transferred to the testing room and acclimated for 30 min before the test session. During the testing period, lighting in the room was turned off. The apparatus was cleaned with quatricide before and between runs. Total movements (distance travelled and velocity) in the arena were quantified using an automated infrared (IR) detection system (EthoVision XT, Noldus). The tracking software plotted heat maps for each mouse, which was averaged to create representative heat maps for each genotype. Raw data were extracted and analysed using Microsoft Excel.

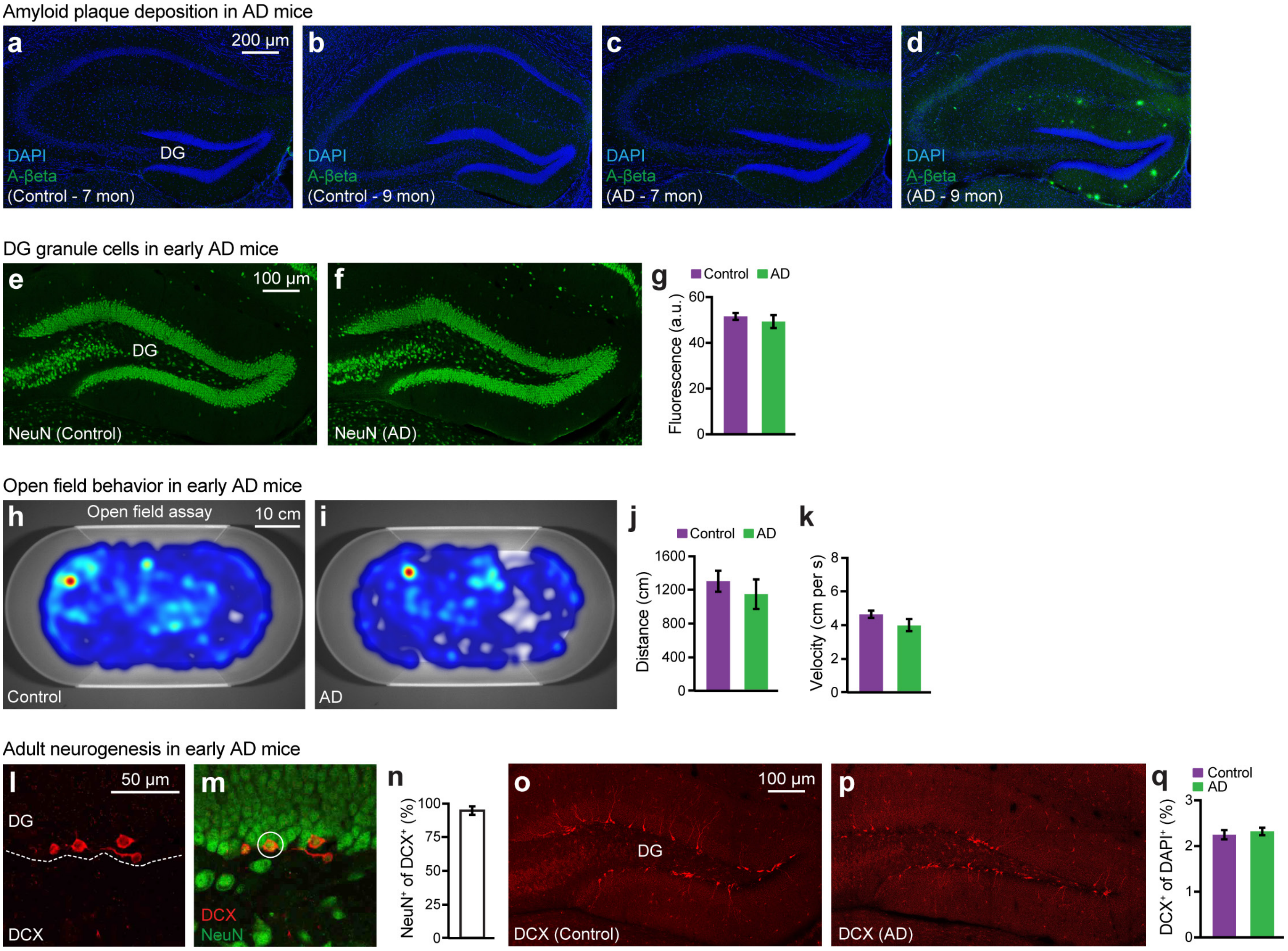
**Engram activation.** For light-induced freezing behaviour, a context distinct from the CFC training chamber (context A) was used. These were 30 × 25 × 33 cm chambers with perspex floors, square ceilings, white lighting, and scented with 0.25% benzaldehyde. Chamber ceilings were customized to hold a rotary joint (Doric Lenses) connected to two 0.32-m patch cords. All mice had patch cords fitted to the optic fibre implant before testing. Two mice were run simultaneously in two identical chambers. ChR2 was stimulated at 20 Hz (15 ms pulse width) using a 473 nm laser (10–15 mW), for the designated epochs. Testing sessions were 12 min in duration, consisting of four 3 min epochs, with the first and third as light-off epochs, and the second and fourth as light-on epochs. At the end of 12 min, the mouse was detached and returned to its home cage. Floors of chambers were cleaned with quatricide before and between runs.

***In vivo* optical LTP.** One day after CFC training and engram labelling (DG plus PP terminals) in control and early AD groups, mice were placed in an open field arena (52 × 26 cm) after patch cords were fitted to the fibre implants. After a 15 min acclimatization period, mice with oChIEF$^+$ PP engram terminals in the DG received the optical LTP[23] protocol (100 blue light pulses of 2 ms each at a frequency of 100 Hz, repeated 5 times every 3 min). This *in vivo* protocol was repeated 10 times over a 3 h duration. After induction, mice remained in the arena for an additional 15 min before returning to their home cage. To apply optical LTP to a large portion of excitatory MEC neurons, an AAV virus expressing oChIEF-tdTomato under the CaMKII promoter, rather than a c-Fos-tTA/TRE virus (that is, engram labelling), was used. For protein synthesis inhibition experiments, immediately after the *in vivo* LTP induction protocol mice received 75 mg kg$^{-1}$ anisomycin (Aniso) or an equivalent volume of saline intraperitoneally. Mice were then returned to their home cages. An hour later, a second injection of Aniso or saline was delivered.

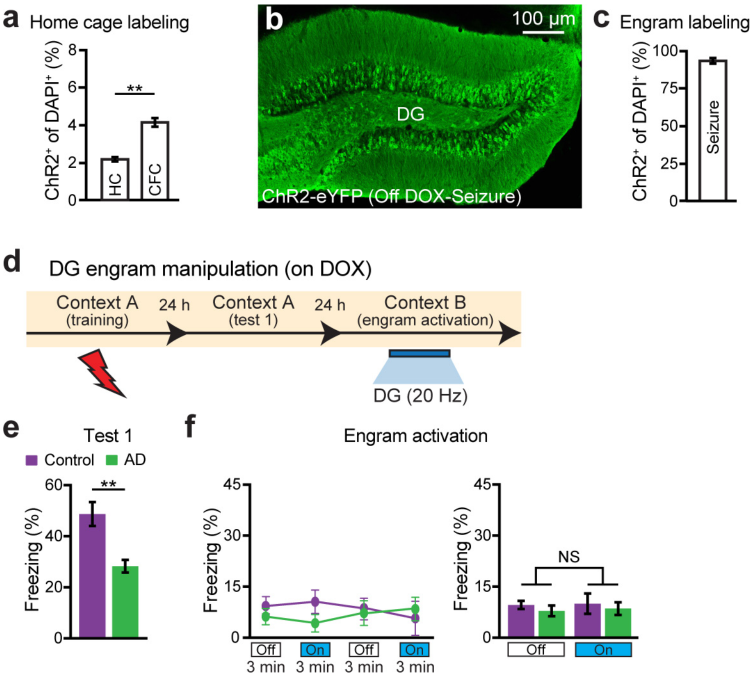
**Inhibitory avoidance.** A 30 × 28 × 34 cm unscented chamber with transparent square ceilings and intermediate lighting was used. The chamber consisted of two sections, one with grid flooring and the other with a white light platform. During the conditioning session (1 min), mice were placed on the light platform, which is the less preferred section of the chamber (relative to the grid section). Once mice entered the grid section of the chamber (all four feet), 0.80 mA shocks of 2 s duration were delivered. On average, each mouse received 2–3 shocks per training session. After 1 min, mice were returned to their home cage. The next day, latency to enter the grid section of the chamber as well as total time on the light platform was measured (3 min test).

**Novel object location.** Spatial memory was measured in a white plastic chamber (28 × 28 cm) that had patterns (series of parallel lines or circles) on opposite walls. The apparatus was unscented and intermediate lighting was used. All mice were transferred to the behavioural room and acclimated for 30 min before the training session. On day 1, mice were allowed to explore the chamber with patterns for 15 min. On days 2 and 3, mice were introduced into the chamber that had an object (7-cm-tall glass flask filled with metal beads) placed adjacent to either patterned wall. The position of the object was counter-balanced within each genotype. On day 4, mice were placed into the chamber with the object either in the same position as the previous exposure (familiar) or at a novel location based on wall patterning. Frequency of visits to the familiar and novel object locations was quantified using an automated detection system (EthoVision XT, Noldus). Total time exploring the object was also measured (nose within 1.5 cm of object). The tracking software plotted heat maps based on exploration time, which was averaged to create representative heat maps for each genotype. Raw data were extracted and analysed using Microsoft Excel.

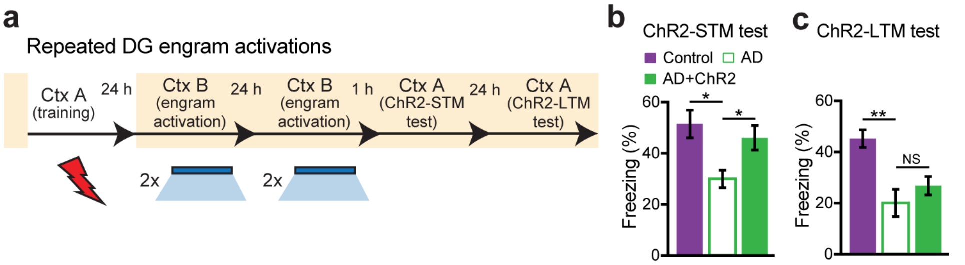
28. Reijmers, L. G., Perkins, B. L., Matsuo, N. & Mayford, M. Localization of a stable neural correlate of associative memory. *Science* **317,** 1230–1233 (2007).
29. Urlinger, S. *et al.* Exploring the sequence space for tetracycline-dependent transcriptional activators: novel mutations yield expanded range and sensitivity. *Proc. Natl Acad. Sci. USA* **97,** 7963–7968 (2000).



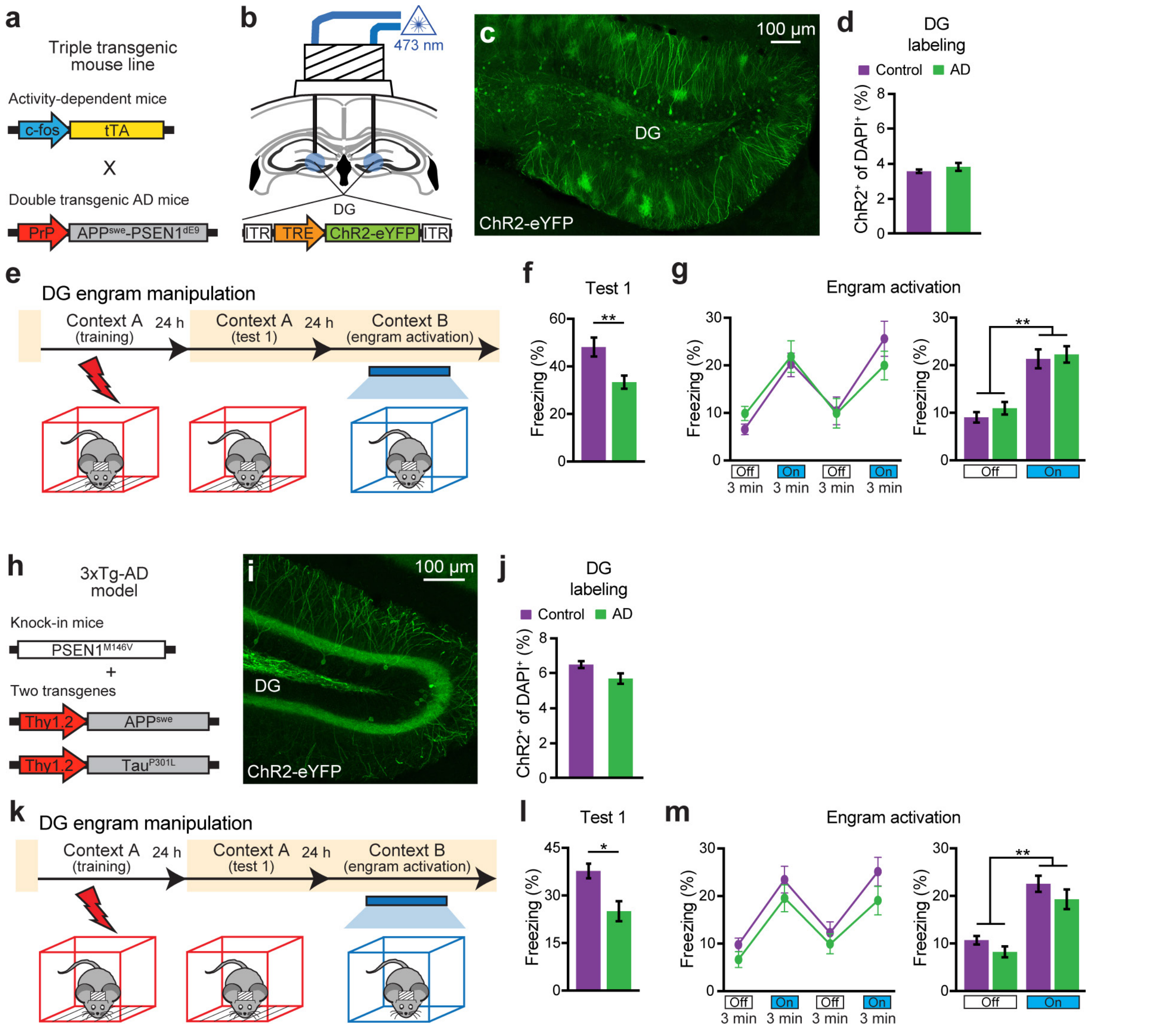
**Extended Data Figure 1 | Characterization of 7-month-old early AD mice. a–d**, Images showing hippocampal A$\beta^+$ plaques lacking in control mice (**a**, **b**) and 7-month-old AD mice (**c**), which showed an age-dependent increase in 9-month-old AD mice (**d**). **e**, **f**, Images showing neuronal nuclei (NeuN) staining of DG granule cells in control (**e**) and 7-month-old AD (**f**) mice. **g**, NeuN$^+$ fluorescence intensity of the granule cell layer from control and AD sections shown in **e**, **f** ($n = 8$ mice per group). **h**, **i**, Heat maps showing exploratory behaviour in an open field arena from control (**h**) and 7-month-old AD (**i**) mice. **j**, **k**, Distance travelled (**j**) and velocity (**k**) did not differ between control and AD groups ($n = 9$ mice per group). **l**, **m**, Images showing adult newborn neurons (DCX$^+$) in DG sections from control mice (**l**) that are double positive for NeuN (**m**). **n**, Percentage of NeuN$^+$ cells among DCX$^+$ cells ($n = 3$ mice). **o**, **p**, Images showing DCX$^+$ neurons in DG sections from control (**o**) and AD (**p**) groups ($n = 4$ mice per group). **q**, DCX$^+$ cell counts from control and AD mice. Data are presented as mean ± s.e.m.



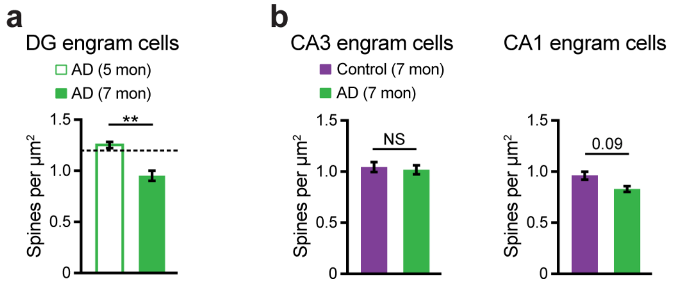
**Extended Data Figure 2 | Labelling and engram activation of early AD mice on DOX.** **a**, Mice are taken off DOX for 24 h in the home cage (HC) and subsequently trained in CFC. DG sections ($n = 3$ mice per group) revealed 2.05% ChR2–eYFP labelling in the home cage, consistent with the previously established engram tagging strategy[11]. **b**, Mice were injected with a virus cocktail of AAV$_9$-c-Fos-tTA and AAV$_9$-TRE-ChR2-eYFP. After 1 day off DOX, kainic acid was used to induce seizures. Image showing efficient labelling throughout the DG. **c**, ChR2–eYFP cell counts from DG sections shown in **b** ($n = 3$ mice). **d**, Behavioural schedule for optogenetic activation of DG engram cells. **e**, Memory recall 1 day after training (test 1) showed less freezing of AD mice compared with control mice ($n = 8$ mice per group). **f**, Engram activation with blue light stimulation (left). Average freezing for the two light-off and light-on epochs (right). Statistical comparisons are performed using unpaired *t*-tests; $^{**}P < 0.01$. Data are presented as mean ± s.e.m.



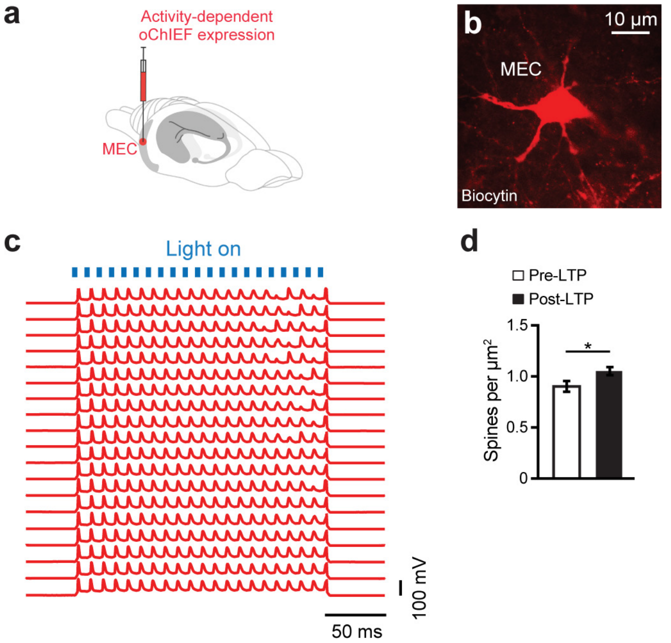
**Extended Data Figure 3 | Chronic DG engram activation in early AD mice did not rescue long-term memory.** **a**, Behavioural schedule for repeated DG engram activation experiment. Ctx, context. **b**, AD mice in which a DG memory engram was reactivated twice a day for 2 days (AD + ChR2) showed increased STM freezing levels compared with memory recall before engram reactivation (ChR2-STM test, *n* = 9 mice per group). **c**, Memory recall 1 day after repeated DG engram activations (ChR2-LTM test). NS, not significant. Statistical comparisons are performed using unpaired *t*-tests; *$P < 0.05$, **$P < 0.01$. Data are presented as mean ± s.e.m.



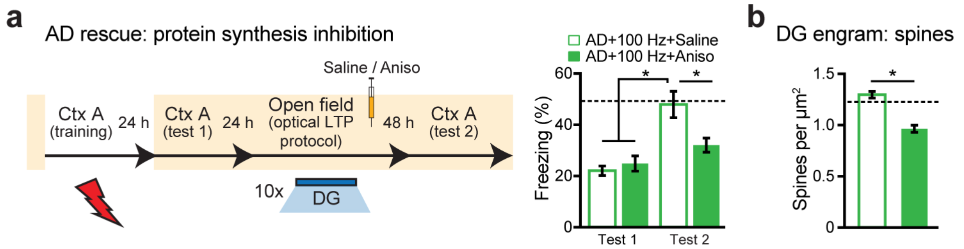
**Extended Data Figure 4 | Engram activation restores fear memory in triple-transgenic and PS1/APP/tau models of early AD. a**, Triple-transgenic mouse line obtained by mating *c-Fos-tTA* transgenic mice[11,28] with double-transgenic APP/PS1 AD mice[10]. These mice combined with a DOX-sensitive AAV virus permits memory engram labelling in early AD. **b**, Triple-transgenic mice were injected with AAV$_9$-TRE-ChR2-eYFP and implanted with an optic fibre targeting the DG. **c**, Image showing DG engram cells of triple-transgenic mice 24 h after CFC. **d**, ChR2–eYFP cell counts from control and triple-transgenic AD mice ($n = 5$ mice per group). **e**, Behavioural schedule for engram activation. **f**, Memory recall 1 day after training (test 1) showed less freezing of triple-transgenic AD mice compared with control mice ($n = 10$ mice per group). **g**, Engram activation with blue light stimulation (left). Average freezing for the two light-off and light-on epochs (right). **h**, Triple-transgenic AD model (3×Tg-AD) as previously reported[18]. A cocktail of AAV$_9$-c-Fos-tTA and AAV$_9$-TRE-ChR2-eYFP viruses were used to label memory engrams in 3×Tg-AD mice. **i**, Image showing memory engram cells in the DG of 3×Tg-AD mice 24 h after CFC. **j**, ChR2–eYFP cell counts from DG sections of control and 3×Tg-AD mice ($n = 4$ mice per group). **k**, Behavioural schedule for engram activation. **l**, Memory recall 1 day after training (test 1) showed less freezing of 3×Tg-AD mice compared with control mice ($n = 9$ mice per group). **m**, Engram activation with blue light stimulation (left). Average freezing for the two light-off and light-on epochs (right). Statistical comparisons are performed using unpaired *t*-tests; $^{*}P < 0.05$, $^{**}P < 0.01$. Data are presented as mean ± s.e.m.



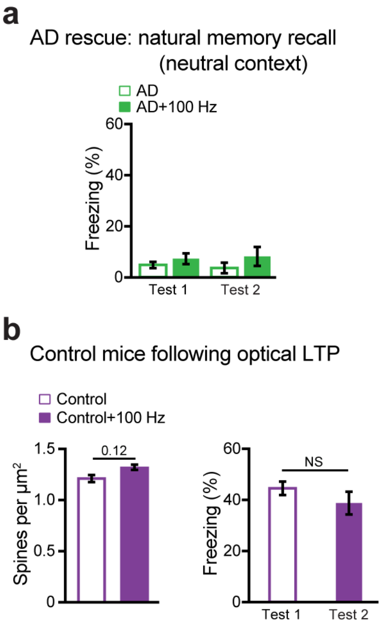
**Extended Data Figure 5 | Dendritic spines of engram cells in 7-month-old early AD mice. a**, Average dendritic spine density of DG engram cells showed an age-dependent decrease in 7-month-old APP/PS1 AD mice ($n = 7{,}032$ spines) as compared to 5-month-old AD mice ($n = 4{,}577$ spines, $n = 4$ mice per group). Dashed line represents spine density of control mice (1.21). **b**, Left, average dendritic spine density of CA3 engram cells in control ($n = 5{,}123$ spines) and AD mice ($n = 6{,}019$ spines, $n = 3$ mice per group). Right, average dendritic spine density of CA1 engram cells in control ($n = 9{,}120$ spines) and AD mice ($n = 7{,}988$ spines, $n = 5$ mice per group). NS, not significant. Statistical comparisons are performed using unpaired *t*-tests; $^{**}P < 0.01$. Data are presented as mean ± s.e.m.



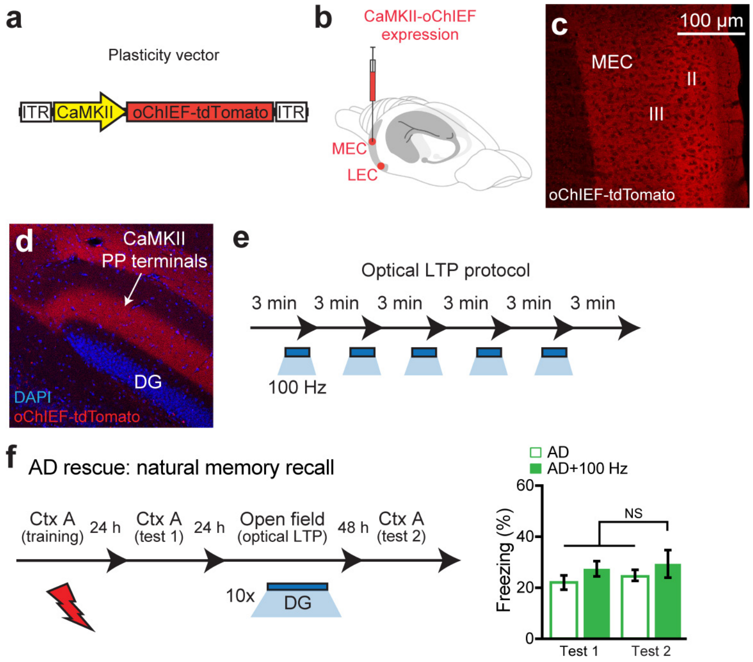
**Extended Data Figure 6 | High-fidelity responses of oChIEF$^+$ cells and dendritic spines of DG engram cells after *in vitro* optical LTP.** **a**, EC cells were injected with a virus cocktail containing AAV$_9$-TRE-oChIEF-tdTomato for activity-dependent labelling. **b**, Image showing a biocytin-filled oChIEF$^+$ stellate cell in the EC. **c**, 100 Hz (2-ms pulse width) stimulation of an oChIEF$^+$ cell across 20 consecutive trials. Spiking responses exhibit high fidelity. **d**, Average dendritic spine density of biocytin-filled DG cells showed an increase after optical LTP induction *in vitro* ($n = 1,452$ spines, $n = 6$ cells). Statistical comparisons are performed using unpaired $t$-tests; $^*P < 0.05$. Data are presented as mean ± s.e.m.



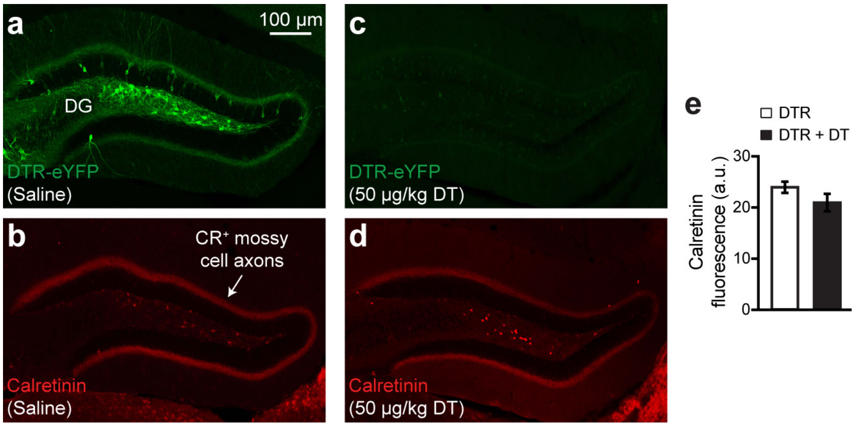
**Extended Data Figure 7 | Behavioural rescue and spine restoration by optical LTP is protein-synthesis dependent. a**, Modified behavioural schedule for long-term rescue of memory recall in AD mice in the presence of saline or anisomycin (left). Memory recall 2 days after LTP induction followed by drug administration showed less freezing of AD mice treated with anisomycin (AD + 100 Hz + Aniso) compared with saline-treated AD mice (AD + 100 Hz + saline, $n = 9$ mice per group; right). Dashed line represents freezing level of control mice (48.53). Ctx, context. **b**, Average dendritic spine density in early AD mice treated with anisomycin after LTP induction ($n = 4{,}810$ spines) was decreased compared with saline-treated AD mice ($n = 6{,}242$ spines, $n = 4$ mice per group). Dashed line represents spine density of control mice (1.21). Statistical comparisons are performed using unpaired $t$-tests; $*P < 0.05$. Data are presented as mean ± s.e.m.



**Extended Data Figure 8 | Rescued early AD mouse behaviour in a neutral context and control mouse behaviour after *in vivo* optical LTP.** **a**, After the long-term rescue of memory recall in AD mice (test 2; Fig. 3m), animals were placed in an untrained neutral context to measure generalization ($n = 10$ mice per group). Rescued AD mice (AD + 100 Hz) did not display freezing behaviour. **b**, Left, average dendritic spine density of DG engram cells from control mice remained unchanged after optical LTP induction *in vivo* (control + 100 Hz, $n = 4{,}211$ spines, $n = 3$ mice; control data from Fig. 2c). Right, the behavioural rescue protocol applied to early AD mice (Fig. 3m) was tested in age-matched control mice ($n = 9$ mice per group). Similar freezing levels were observed after optical LTP (test 2) as compared to memory recall before the 100 Hz protocol (test 1). NS, not significant. Statistical comparisons are performed using unpaired *t*-tests. Data are presented as mean ± s.e.m.



**Extended Data Figure 9 | Optical LTP using a CaMKII-oChIEF virus did not rescue memory in early AD mice. a**, AAV virus expressing oChIEF-tdTomato under a CaMKII promoter. **b**, CaMKII-oChIEF virus injected into MEC and LEC. **c**, **d**, Images showing tdTomato labelling in a large portion of excitatory MEC neurons (**c**) as well as the PP terminals in the DG (**d**). **e**, *In vivo* optical LTP protocol[23]. **f**, Behavioural schedule for long-term rescue of memory recall in AD mice (left). In contrast to the engram-specific strategy, long-term memory could not be rescued by stimulating a large portion of excitatory PP terminals in the DG (right; *n* = 9 mice per group). NS, not significant. Statistical comparisons are performed using unpaired *t*-tests. Data are presented as mean ± s.e.m.



**Extended Data Figure 10 | Normal DG mossy cell density after engram cell ablation. a–d**, Images showing DG engram cells after saline treatment (**a**) and the corresponding calretinin positive (CR$^+$) mossy cell axons (**b**). DTR–eYFP engram cell labelling after DT treatment (**c**) and the respective CR$^+$ mossy cell axons (**d**). **e**, CR$^+$ fluorescence intensity of mossy cell axons from saline- and DT-treated DG sections shown in **a**–**d** ($n = 8$ mice per group). Data are presented as mean ± s.e.m.

doi:10.1038/nature16985

# Visualization of immediate immune responses to pioneer metastatic cells in the lung

Mark B. Headley[1], Adriaan Bins[1,2], Alyssa Nip[1], Edward W. Roberts[1], Mark R. Looney[3], Audrey Gerard[1] & Matthew F. Krummel[1]

**Lung metastasis is the lethal determinant in many cancers[1,2] and a number of lines of evidence point to monocytes and macrophages having key roles in its development[3–5]. Yet little is known about the immediate fate of incoming tumour cells as they colonize this tissue, and even less known about how they make first contact with the immune system. Primary tumours liberate circulating tumour cells (CTCs) into the blood and we have developed a stable intravital two-photon lung imaging model in mice[6] for direct observation of the arrival of CTCs and subsequent host interaction. Here we show dynamic generation of tumour microparticles in shear flow in the capillaries within minutes of CTC entry. Rather than dispersing under flow, many of these microparticles remain attached to the lung vasculature or independently migrate along the inner walls of vessels. Using fluorescent lineage reporters and flow cytometry, we observed 'waves' of distinct myeloid cell subsets that load differentially and sequentially with this CTC-derived material. Many of these tumour-ingesting myeloid cells collectively accumulated in the lung interstitium along with the successful metastatic cells and, as previously understood, promote the development of successful metastases from surviving tumour cells[3]. Although the numbers of these cells rise globally in the lung with metastatic exposure and ingesting myeloid cells undergo phenotypic changes associated with microparticle ingestion, a consistently sparse population of resident conventional dendritic cells, among the last cells to interact with CTCs, confer anti-metastatic protection. This work reveals that CTC fragmentation generates immune-interacting intermediates, and defines a competitive relationship between phagocyte populations for tumour loading during metastatic cell seeding.**

Primary tumours induce distal accumulation of immune cells in the lung that promotes metastasis[4,5]. B16F10 subcutaneous tumours, expressing ZsGreen (hereafter referred to as B16ZsGreen), resulted in CD45$^+$ZsGreen$^+$ (immune) cells in the lung before the appearance of micrometastases (Fig. 1a, b). These cells contained vesicular ZsGreen$^+$ puncta, suggesting ingestion of tumour fragments (Fig. 1c). An experimental metastasis model, intravenous (i.v.) injection of B16ZsGreen cells, revealed similar loading of intracellular vesicles in CD45$^+$ cells over 24 h (Extended Data Fig. 1a–c). Notably, CD45$^+$ZsGreen$^+$ cells increased rapidly within 4 h after injection, and exceeded the initial frequency of the B16ZsGreen cells (Extended Data Fig. 1c), suggesting a cell-fragment origin of ingested tumour material. These CD45$^+$ZsGreen$^+$ cells again had puncta of ingested tumour material (Extended Data Fig. 1d).

Previous work has established roles for primary tumour-derived exosomes ($\leq$200 nm in diameter) in the lung pre-metastatic niche[7]. However, the size of inclusions within sorted CD45$^+$ZsGreen$^+$ cells (diameters often $\geq$1 μm (Fig. 1c and Extended Data Fig. 1d)) suggested that this might occur via a distinct mechanism. As CTCs are detectable in the blood of patients with metastatic disease[8], we sought to visualize their fate upon entering the lung vasculature. To image the arrival of injected B16ZsGreen metastatic cells in lung capillaries, we updated our published method for lung intravital microscopy (LIVM)[6] using a novel intercostal window (Extended Data Fig. 2a–e) allowing for stable imaging of $\leq$12 h (Extended Data Fig. 2f). Figure 1d and Supplementary Video 1 show that within seconds of arrival, the incoming CTCs became lodged in capillaries and began to shed microscale blebs (microparticles) into the vasculature.

Microparticle generation was observed repeatedly over at least 8 h after injection (Fig. 1e and Supplementary Video 2). These blebs had an average diameter of 5 μm, with a range of 0.5 to >25 μm (Fig. 1f), markedly larger than exosomes and larger than most previously reported microparticles[9]. Arrested B16ZsGreen cells were also observed to undergo membrane protrusion and retraction (Supplementary Video 2 and Extended Data Fig. 3a). Over time, surviving tumour cells decreased in size and the protrusive activity reduced (Extended Data Fig. 3a–c and Supplementary Video 3). In most instances of microparticle release, the nucleus of the parent cell maintained integrity and the blebs themselves retained the cytoplasmic fluorophore, consistent with non-apoptotic blebbing[10]. We also observed cells undergoing lysis (Supplementary Video 4), distinguished from apoptosis by a lack of blebbing and rapid loss of the cytoplasmic fluorophore. On the basis of these criteria, we adopted the terminology for parental nucleated cells as 'karyoplasts' and the blebs as 'cytoplasts'.

Previous lung explant and slice imaging studies did not reveal the generation of cytoplasts during metastasis[11,12]; we hypothesized that their formation was dependent on shear forces in intact lungs. We compared the behaviour of tumour cells in LIVM versus slice to test this[13]. In the non-shear system, the production of cytoplasts was approximately fourfold reduced (Fig. 1g). Further, cytoplast production was unaltered in the presence of the apoptosis inhibitor Z-VAD (Extended Data Fig. 4a, b). This demonstrates that cytoplast formation is driven by physical forces rather than a programmed cell death mechanism. In LIVM, many microparticles also exhibited autonomous motility, with spontaneous arrest and adherence on vascular walls (Supplementary Video 5) and migration against vascular flow (Extended Data Fig. 4c–e and Supplementary Video 6). The speed of these particles was around 10 μm min$^{-1}$ in the absence of vascular flow, and an order of magnitude faster in LIVM (Fig. 1h).

Having observed the generation process *in vivo*, we developed a flow cytometric assay to study their formation and fate. Using Hoechst-staining to accurately differentiate karyoplasts from cytoplasts (confirmed by confocal imaging), we found that cytoplasts became the dominant tumour-cell-derived species within the lung within 15 min (Fig. 1i, j and Extended Data Fig. 3d). Cytoplasts contained mitochondria (Fig. 1k, l and Extended Data Fig. 3e), consistent with retention of metabolic potential and motility. The production of these particles was not unique to *in vitro* passaged B16F10 cells, as robust cytoplast generation was seen with B16ZsGreen cells

[1]Department of Pathology, University of California, San Francisco, 513 Parnassus Ave, HSW512, San Francisco, California 94143-0511, USA. [2]Department of Medical Oncology, Academic Medical Center Amsterdam, Meibergdreef, 91105AZ Amsterdam, The Netherlands. [3]Departments of Medicine and Laboratory Medicine, University of California, San Francisco, 513 Parnassus Avenue, HSW512, California 94143-0511, USA.

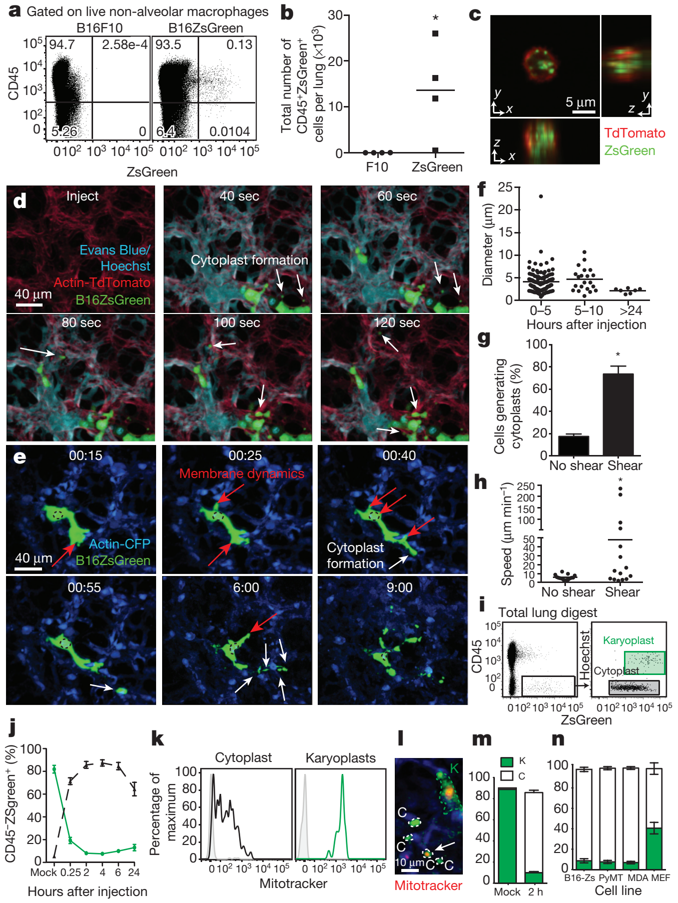


**Figure 1 | Intravital imaging of the first hours of lung seeding by B16 melanoma. a**, Representative plots of CD45$^+$ZSgreen$^+$ cells in lungs of mice bearing 2-week primary B16 melanoma tumours with or without ZsGreen expression. Alveolar macrophages were excluded, as auto-fluorescent signal interfered with ZsGreen discrimination. **b**, Absolute number of CD45$^+$ZsGreen$^+$ and CD45$^-$ZsGreen$^+$ cells in lungs from **a**. *$P = 0.0421$ by unpaired *t*-test. **c**, Confocal imaging of sorted CD45$^+$ZsGreen$^+$ cells from lungs of mTmG mouse (with ubiquitous expression of membrane-bound TdTomato) bearing tumours as per **a**. Red, TdTomato; green, ZsGreen. **d**, **e**, LIVM after i.v. injection of Hoechst-labelled B16ZsGreen cells into mTmG mice with Evan's Blue labelling vascular flow (see also Supplementary Video 1) (**d**). LIVM Hoechst-labelled B16ZsGreen cells from 15 min to 9 h after i.v. injection into actin–CFP recipient (see also Supplementary Video 2) (**e**). White arrows highlight the formation of cytoplasmic blebs (cytoplasts) and red arrows highlight regions of membrane activity (extension or retraction). Representative of at least 10 mice. **f**, Cytoplast diameter at 0–5 h, 5–10 h and >24 h after injection. **g**, Percent cytoplast producing B16ZsGreen cells from non-shear (slice imaging) or shear (LIVM) (15 cells per group from 3 mice; *$P = 0.002$ by unpaired *t*-test). **h**, Cytoplast speed in non-shear or shear conditions as **f** (15 cytoplasts per group in 3 mice; *$P = 0.036$, unpaired *t*-test). **i**, Gating strategy for karyoplast and cytoplast discrimination within lung single-cell suspension. **j**, *In vivo* mean cytoplast and karyoplast frequency in lung ($n = 6$ per group). **k**, Mitotracker staining of cytoplasts and karyoplasts. **l**, LIVM of mitotracker labelled B16ZsGreen cells. **m**, *In vivo* cytoplast frequency in lung 2 h after injection of primary-isolated B16ZsGreen cells or mock injected. Green, karyoplasts; white, cytoplasts. $n = 4$ per group. **n**, *In vivo* cytoplast frequency 2 h after injection of B16ZsGreen, murine breast tumour (PyMT-B), human breast tumour (MDA-MB-231), and non-transformed primary mouse embryonic fibroblasts ($n = 3$ per group from 2 experiments). Horizontal bars represent means, error bars are s.d.

isolated directly from primary subdermal tumours and re-injected i.v. (Fig. 1m), a mouse breast tumour line (PyMT-B), human MDA-MB231 breast tumour cells, and non-transformed mouse embryonic fibroblasts (Fig. 1n).

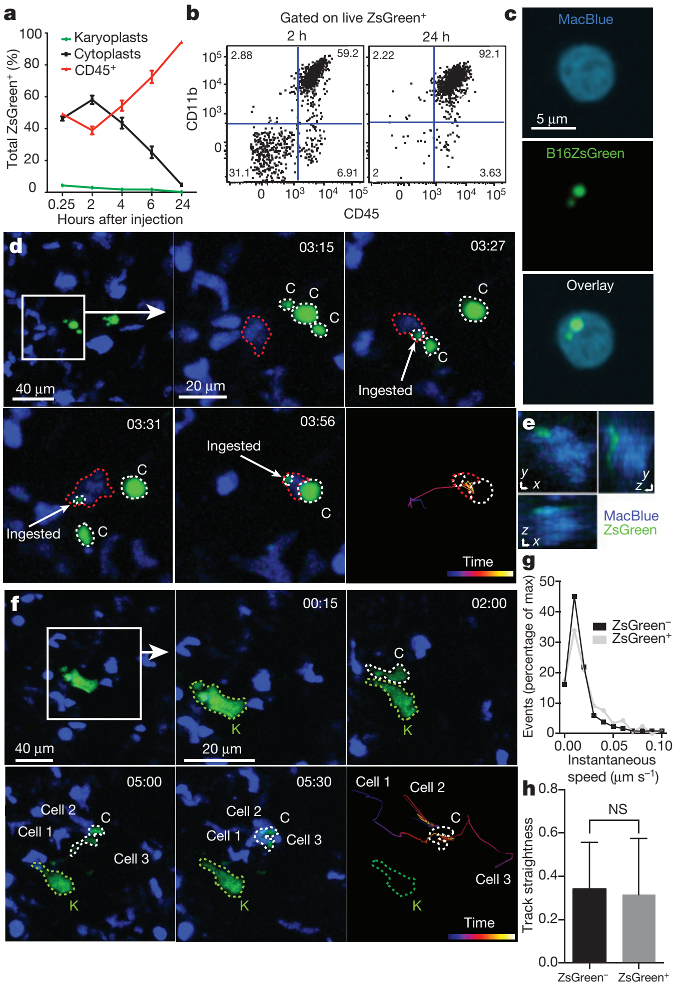


**Figure 2 | Encounter and uptake of tumour-derived cytoplasts by lung myeloid cells. a**, Frequencies of cytoplasts, karyoplasts, and CD45$^+$ZsGreen$^+$ cells in the lung over 24 h ($n = 6$ per group). **b**, Flow cytometry of CD11b$^+$CD45$^+$ cells in total ZsGreen$^+$ population 2 and 24 h after injection. **c**, Confocal imaging of CD45$^+$ZsGreen$^+$ cells from MacBlue mice 24 h after injection. **d**, LIVM of cytoplast phagocytosis by a CFP$^+$ myeloid cell (see also Supplementary Video 7). Green, karyoplasts; white, cytoplasts; red, cytoplast ingesting cell. Rightmost panel shows tracking data for a subset of cells. Representative of 5 mice. **e**, *xy*, *yz* and *xz* renders of CFP$^+$ myeloic cell from **d**. **f**, LIVM of CFP$^+$ cell targeting cytoplast (see also Supplementary Video 8). Colours and tracking as in **d**. Tracked cells labelled as cell 1, 2 or 3 in **f** for comparison with tracks. Representative of 5 mice. **g**, Speed of cytoplast-ingesting and non-ingesting cells. **h**, Track straightness of cytoplast-ingesting and non-ingesting cells. Data from 75 non-ingesting and 10 ingesting cells from 4 mice, error bars are s.d.

The number of cytoplasts peaked by 4 h, with few detectable by 24 h. Meanwhile, karyoplast number declined according to a one-phase exponential decay over the first day with a half-life ($t_{1/2}$) of 6.3 h (Fig. 2a). Comparison of the frequency of immune-associated ZsGreen$^+$ events with that of free cytoplasts and karyoplasts revealed a strong reciprocal relationship between the rise in ZsGreen$^+$ immune cells and loss of free cytoplasts (Fig. 2a). This suggested that cytoplasts represent the source of ingested tumour material in pre-metastatic lungs. Myeloid cells were implicated as the primary phagocytes on the basis of CD11b expression (Fig. 2b). Assessment of the myeloid lineage reporter MacBlue (eCFP (enhanced cyan fluorescent protein) expression driven by a modified *CFMS* promoter cassette, labelling monocytes and monocyte-derived cells, as well as a small portion of neutrophils)[14] further revealed CFP$^+$ZsGreen$^+$ cells containing ingested tumour fragments (≥1 μm diameter), consistent with cytoplast

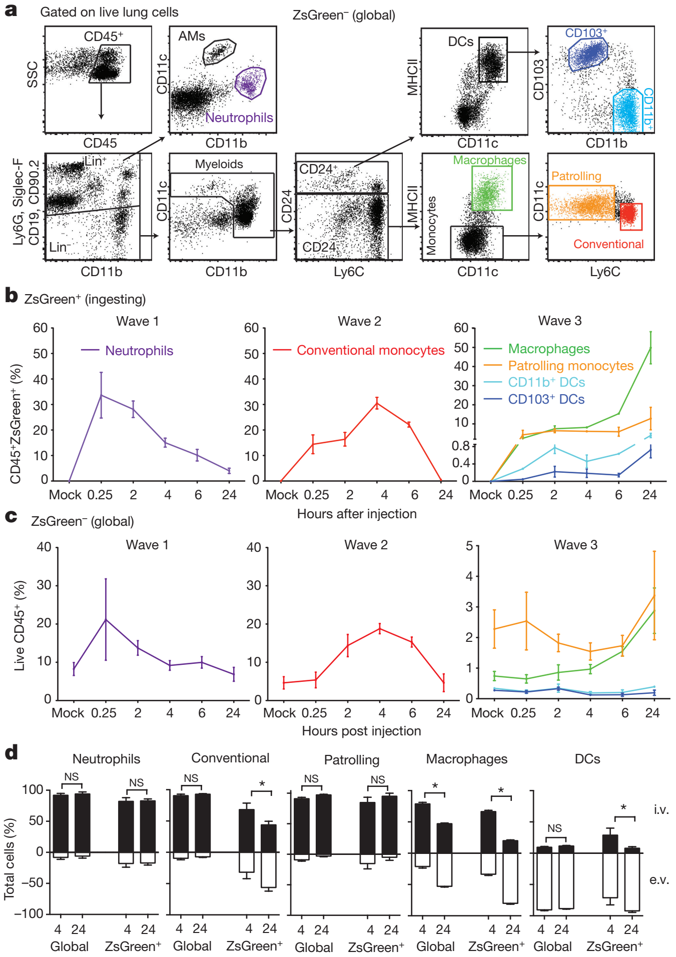


**Figure 3 | Discrete waves of cytoplast loaded myeloid cells define the early metastatic niche. a**, Gating strategy for total lung myeloid populations. **b**, Frequency of tumour-ingesting myeloid cells in the lung over 24 h following i.v. injection with B16ZsGreen ($n = 6$). **c**, Frequency of myeloid cells in total lung cells over 24 h following i.v. injection with B16ZsGreen ($n = 6$). **d**, Frequency of intravascular versus extravascular myeloid populations at 4 or 24 h following i.v. injection of B16ZsGreen cells ($n = 6$ per group; $^*P < 0.05$, two-way ANOVA with multiple comparison between row and column means). Error bars are s.d.

uptake (Fig. 2c). Importantly, using LIVM, we were able to observe CFP$^+$ cells directly ingesting cytoplasts (Fig. 2d, e and Supplementary Video 7). Additionally, we occasionally observed swarming by CFP$^+$ cells following release of a cytoplast from a parental karyoplast with these cells ignoring the neighbouring viable karyoplast (Fig. 2f and Supplementary Video 8). A bias of phagocytes towards cytoplasts was not universal and CFP$^+$ cells were also frequently found in direct interaction with karyoplasts (Supplementary Video 9), although in hundreds of hours of imaging, we never observed phagocytosis of an intact karyoplast. Evaluation of the behavioural characteristics of cytoplast-ingesting versus non-ingesting myeloid cells revealed no clear differences in either instantaneous speed nor path straightness, although the latter would be heavily dictated by the vasculature (Fig. 2g, h). Taken together, these data support the hypothesis that pioneer CTCs generate cytoplasts within the lung vasculature early on arrival in the lung, leading to loading of local phagocytes.

We next characterized the identity of the tumour-ingesting myeloid cells in the early metastatic niche using flow cytometry (Fig. 3a and Extended Data Fig. 5a). Our analysis included specific gating of alveolar macrophages (Siglec-F$^+$, CD11c$^+$, CD11b$^{\text{low}}$), neutrophils (Ly6G$^+$, Ly6C$^+$, CD11b$^+$), conventional monocytes (Ly6C$^+$, CD11b$^+$, MHCII$^{+/-}$, Ly6G$^-$, CD11c$^-$, CD24$^-$), patrolling monocytes (CD11b$^+$, CD11c$^{\text{mid}}$, Siglec-F$^-$, Ly6G$^-$, MHCII$^-$, CD24$^-$, Ly6C$^{\text{low}/-}$), non-alveolar macrophages (CD11b$^+$, CD11c$^+$, MHCII$^+$, CD24$^-$, Siglec-F$^-$, Ly6G$^-$) and two populations of lung-resident conventional dendritic cells (cDCs): CD103$^+$ and CD11b$^+$ (CD24$^+$, CD11c$^+$, MHCII$^{\text{hi}}$, Ly6C$^-$, CD103$^+$ or CD11b$^+$). Both monocyte populations expressed CD115, and macrophages, but not cDCs, expressed CD64 and F4/80, supporting the definitions of these populations (data not shown). Using this strategy, we examined lungs across a 24 h time course, gating on ZsGreen$^-$ (Fig. 3a) or ZsGreen$^+$ (Extended Data Fig. 5a) cells to understand the dynamics of total versus cytoplast-loaded myeloid populations. This revealed a progressive wavelike shift in the identity of myeloid cells that contained CTC-derived material (ZsGreen$^+$) (Fig. 3b, c). The first wave was dominated by neutrophil uptake, occurring within 15 min of injection, peaking around 30 min and returning to near-baseline by 24 h. Wave 2 was dominated by loading of conventional monocytes with a peak at 4 h followed by a return to baseline by 24 h. Finally, wave 3 established between 6 and 24 h comprised of non-alveolar macrophages, patrolling monocytes and DCs. Among those, non-alveolar macrophages were the most numerous, representing greater than 60% of total CD45$^+$ZsGreen$^+$ cells at 24 h, followed by patrolling monocytes and both CD103$^+$ DCs and CD11b$^+$ DCs at much lower frequencies (Fig. 3b). These waves were largely mirrored in global (ZsGreen$^-$) populations with the notable exception of the DC populations (Fig. 3c), consistent with uptake frequency being largely a function of cellular frequency.

We next sought to determine how cytoplast-ingesting cells behaved with respect to extravasation. We used i.v. injection of labelled anti-CD45 antibody[15] to localize ZsGreen$^+$ populations relative to the vasculature at 4 and 24 h after injection. Consistent with previous data[3], non-alveolar macrophages extravasated over this time period (Fig. 3d and Extended Data Fig. 5b–d). Interestingly, a portion of ZsGreen$^+$ conventional monocytes also extravasated, consistent with recent reports that this population can exit circulation without differentiation into macrophages[16]. In contrast, tumour-loaded patrolling monocytes and neutrophils remained intravascular throughout. The DC populations that gathered tumour material remained largely extravascular throughout, with the exception of a very small but consistent population of ZsGreen$^+$ DCs, at 4 h, that stained for i.v. CD45, which may reflect intravascular sampling by these DC populations (Fig. 3d).

It is likely that some elements of this wave structure are influenced by the bolus nature of the i.v. injection model in contrast to the more continual release of CTCs expected in spontaneous metastasis. Although it is currently difficult to catch cells arriving from a spontaneous tumour, we hypothesize that these waves would exist as a continuum; indeed, the cytoplast-containing cells in lungs of mice bearing primary tumours were dominated by monocytes and macrophages (Extended Data Fig. 5e). Nonetheless, these data suggest that a dichotomy exists in the behaviour of recruited versus lung-resident myeloid populations during metastasis.

Multicolour imaging at 24 h confirmed previous observations[4] that monocyte and monocyte-derived cells (CFP$^+$) and not DCs were dominant in a successful early metastasis (Fig. 4a). Using B16ZsGreen cells, our study extends those observations by showing that these cells contain ingested ZsGreen$^+$ tumour material. This suggests that some originated as cytoplast-ingesting cells (Fig. 4b). Live-imaging of these nascent metastases showed relatively stable and non-motile myeloid cells, although a few underwent directional migration towards micrometastases (Fig. 4c, d). Notably, ZsGreen$^+$ macrophages displayed unique upregulation of a variety of adhesion and chemotactic receptors, relative to ZsGreen$^-$ or steady-state macrophages (Fig. 4e). This provides evidence that tumour ingestion and interaction drives these cells down a distinct activation pathway.

The loaded DC wave at 24 h suggested the potential for an immune-stimulatory fate for some tumour-derived material. Flow-cytometric analysis of antigen-presenting cell (APC) populations

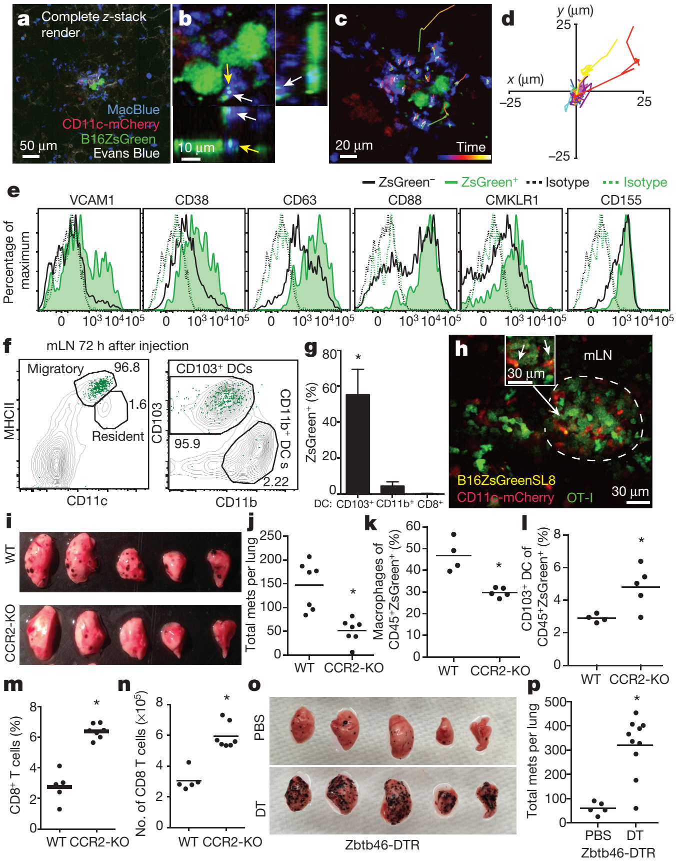


**Figure 4 | Lung-resident dendritic cells inhibit metastasis of B16 melanoma. a**, Maximum intensity projection of LIVM of B16ZsGreen tumour 24 h after injection into CD11c-mCherry MacBlue host (see also Supplementary Video 9). Representative of 5 mice. **b**, High magnification showing tumour-loaded MacBlue$^+$ cells in close interaction with B16ZsGreen cell at 24 h after injection. **c**, Representative image from time-lapse. Track overlays describe motion of tracked MacBlue$^+$ cells over a period of ~1 h. **d**, Overlay of 10 tracks of MacBlue$^+$ cells from **c**. **e**, Staining of ZsGreen$^+$ and ZsGreen$^-$ macrophages 24 h after tumour injection ($n = 4$). **f**, Gating of lung-draining mLN 72 h after injection. Contour lines, total LN; green dots, ZsGreen$^+$ cells. **g**, Frequency of CD8$^+$, CD11b$^+$ and CD103$^+$ DCs in total ZsGreen$^+$ cells in the mLN 72 h after injection. ($n = 6$; *$P < 0.05$, one-way ANOVA with Bonferroni post-hoc test). **h**, Representative image of ZsGreen$^+$CD11c-mCherry$^+$ APCs interacting with GFP-labelled OT-I T cells in a mLN 72 h after injection with B16ZsGreenSL8 (see also Supplementary Video 10). Inset highlights two ZsGreen-bearing CD11c$^+$ cells in close interaction with OT-I T cells ($n = 3$ mice). **i**, Representative images of metastasis-bearing lungs from wild-type (WT) or CCR2-knockout (KO) mice 2 weeks after injection with B16ZsGreen. **j**, Total number of lung metastases (mets) in wild-type and CCR2-knockout mice 2 weeks after injection with B16ZsGreen ($n = 7$ per group; *$P < 0.05$, unpaired $t$-test). **k**, Frequency of ZsGreen$^+$ macrophages in lungs of wild-type and CCR2-knockout mice 24 h after injection with B16ZsGreen ($n = 4$ per group; *$P < 0.05$ by unpaired $t$-test). **l**, Frequency of ZsGreen$^+$ CD103$^+$ cDCs in lungs of wild-type and CCR2-knockout mice 24 h after injection with B16ZsGreen ($n = 4$ per group; *$P < 0.05$, unpaired $t$-test). **m**, **n**, Frequency (**m**) and absolute number (**n**) of CD8$^+$ T cells in lungs of wild-type or CCR20-knockout mice from **k** ($n = 4$ per group; *$P < 0.05$, unpaired $t$-test). **o**, Representative images of lungs from Zbtb46-DTR bone marrow chimaeras treated with phosphate buffered saline (PBS) or diphtheria toxin (DT). Lungs collected 2 weeks after injection with B16ZsGreen. **p**, Total number of lung metastases in PBS- or DT-treated Zbtb46-DTR bone marrow chimaeras 2 weeks after injection with B16ZsGreen ($n = 5$ for PBS and 10 for DT groups, respectively; *$P < 0.05$, unpaired $t$-test). Horizontal bars represent means, error bars are s.d.

in the mediastinal lymph node (mLN) revealed that all ZsGreen$^+$ cells were contained within the migratory CD103$^+$ DC population[17] (Fig. 4f–g). We thus imaged live mLN explants from CD11c-mCherry reporter mice (labelling DC and macrophages through expression of mCherry under control of the *Cd11c* promoter), 72 h after injection with B16F10 cells expressing ZsGreen and the ovalbumin peptide SL8 (B16ZsGreenSL8.) We found clusters of mCherry$^+$ APCs with ZsGreen puncta, clustered with transferred ovalbumin-specific CD8$^+$ (OT-I) T cells. The T cells exhibited a blast-like morphology, suggesting *in vivo* activation of T cells by DCs within the mLN (Fig. 4h and Supplementary Video 10). These data support a model whereby CD103$^+$ DCs acquire tumour material in the lung and subsequently migrate to the mLN to engage with cognate T cells. *In vitro*, CD103$^+$ DCs (isolated from mLN of metastasis-bearing mice at 72 h) were superior in their ability to activate OT-I T cells (Extended Data Fig. 6a–d). In contrast, neither CD11b$^+$ DCs nor CD8$^+$ DCs were observed to contain ZsGreen and were unable to stimulate OT-I T cells *in vitro* (Extended Data Fig. 6a–d).

Previous studies have highlighted the importance of the CCR2–CCL2 axis in monocyte recruitment to seed pro-tumour macrophage populations in the lung[4]. Analysis of CCR2-knockout animals, which recruit significantly fewer monocytes, and consequently macrophages[4], at two weeks after injection with B16ZsGreen, confirmed that CCR2-knockout hosts developed significantly fewer metastases (Fig. 4i, j). By characterizing ZsGreen$^+$ cells at 24 h, we observed a shift in cytoplast loading towards the immunostimulatory CD103$^+$ DCs (Fig. 4k, l). Moreover, in CCR2-knockout animals, we found a substantial increase both in the frequency and number of CD8$^+$ T cells in the lung (Fig. 4m, n). We thus considered whether host-protective CD103$^+$ DCs operate in competition with the pro-tumour macrophages.

To test the idea that metastatic success is specifically opposed by cDCs, we generated Zbtb46-DTR (expression of diphtheria toxin under control of the cDC-specific *Zbtb46* promoter)[18] bone marrow chimaeras. This permitted us to specifically and temporally deplete cDCs. We evaluated the metastatic burden in lungs two weeks after injection with B16ZsGreen; cDC-depleted animals developed sixfold more metastases than the non-depleted controls (Fig. 4o, p). We repeated these experiments in the presence of a subdermal primary tumour, mimicking the endogenous state of tumour metastasis with identical results (Extended Data Fig. 7). These data clearly establish that, despite relative rarity in the tumour interacting pool, lung-resident cDCs have a major role in restricting metastasis.

Previous research into how lung immune responses dictate metastatic success has been largely limited to recruited myeloid populations. This is evidenced by a recent review of the literature[2] that detailed 28 distinct primary articles focused on the function of recruited immune cells in the lung during metastasis and no references defining a role for resident immune cells. Although our analysis of tumour-interacting myeloid cells strongly supports waves of recruited neutrophils, monocytes and macrophages as populations of interest in the early metastatic niche, it also reveals the presence of rare lung-resident cDC subsets within the early interacting pool that mediate potent anti-metastatic effects. A recent study from our lab[19] identified a similarly rare CD103$^+$ cDC population acting locally in primary tumours to stimulate CD8$^+$ T cells providing adaptive protection. This suggests that novel immunotherapeutics designed to enhance these DC populations may provide a single strategy to modulate both primary and metastatic responses. This real-time study, like others of its kind, also highlights the value of direct observation as a means to identify how tumours interact with the colonized tissue, in this case by feeding material from the first pioneer cells into specific subsets of local and recruited myeloid populations.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 19 March 2015; accepted 12 January 2016.**
**Published online 16 March 2016.**

1. Nguyen, D. X., Bos, P. D. & Massague, J. Metastasis: from dissemination to organ-specific colonization. *Nature Rev. Cancer* **9,** 274–284 (2009).
2. Kitamura, T., Qian, B. Z. & Pollard, J. W. Immune cell promotion of metastasis. *Nature Rev. Immunol.* **15,** 73–86 (2015).
3. Qian, B. *et al.* A distinct macrophage population mediates metastatic breast cancer cell extravasation, establishment and growth. *PLoS ONE* **4,** e6562 (2009).
4. Qian, B. Z. *et al.* CCL2 recruits inflammatory monocytes to facilitate breast-tumour metastasis. *Nature* **475,** 222–225 (2011).
5. Kaplan, R. N. *et al.* VEGFR1-positive haematopoietic bone marrow progenitors initiate the pre-metastatic niche. *Nature* **438,** 820–827 (2005).
6. Looney, M. R. *et al.* Stabilized imaging of immune surveillance in the mouse lung. *Nature Methods* **8,** 91–96 (2011).
7. Peinado, H. *et al.* Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET. *Nature Med.* **18,** 883–891 (2012).
8. Joyce, J. A. & Pollard, J. W. Microenvironmental regulation of metastasis. *Nature Rev. Cancer* **9,** 239–252 (2009).
9. Raposo, G. & Stoorvogel, W. Extracellular vesicles: exosomes, microvesicles, and friends. *J. Cell Biol.* **200,** 373–383 (2013).
10. Di Vizio, D. *et al.* Oncosome formation in prostate cancer: association with a region of frequent chromosomal deletion in metastatic disease. *Cancer Res.* **69,** 5601–5609 (2009).
11. Mendoza, A. *et al.* Modeling metastasis biology and therapy in real time in the mouse lung. *J. Clin. Invest.* **120,** 2979–2988 (2010).
12. Al-Mehdi, A. B. *et al.* Intravascular origin of metastasis from the proliferation of endothelium-attached tumor cells: a new model for metastasis. *Nature Med.* **6,** 100–102 (2000).
13. Thornton, E. E. *et al.* Spatiotemporally separated antigen uptake by alveolar dendritic cells and airway presentation to T cells in the lung. *J. Exp. Med.* **209,** 1183–1199 (2012).
14. Ovchinnikov, D. A. *et al.* Expression of Gal4-dependent transgenes in cells of the mononuclear phagocyte system labeled with enhanced cyan fluorescent protein using *Csf1r*-Gal4VP16/UAS-ECFP double-transgenic mice. *J. Leukoc. Biol.* **83,** 430–433 (2008).
15. Anderson, K. G. *et al.* Intravascular staining for discrimination of vascular and tissue leukocytes. *Nature Protocols* **9,** 209–222 (2014).
16. Jakubzick, C. *et al.* Minimal differentiation of classical monocytes as they survey steady-state tissues and transport antigen to lymph nodes. *Immunity* **39,** 599–610 (2013).
17. Idoyaga, J. *et al.* Specialized role of migratory dendritic cells in peripheral tolerance induction. *J. Clin. Invest.* **123,** 844–854 (2013).
18. Meredith, M. M. *et al.* Expression of the zinc finger transcription factor zDC (Zbtb46, Btbd4) defines the classical dendritic cell lineage. *J. Exp. Med.* **209,** 1153–1165 (2012).
19. Broz, M. L. *et al.* Dissecting the tumor myeloid compartment reveals rare activating antigen-presenting cells critical for T cell immunity. *Cancer Cell* **26,** 638–652 (2014).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We thank D. Hume for providing MacBlue mice; Z. Werb for providing MDA-MB231 GFP cells; J. Massegue and S. Abrams for providing PyMT-B cells; J. Cyster for providing us with CCR2-knockout mice. We would also like to thank E. Thornton for her initial work in developing lung intravital imaging. M. Werner for early technical assistance in this project. All members of the Krummel laboratory, BIDC, and M Koch for discussion, support, and guidance while developing this work. This work was supported in part by a Department of Defense post-doctoral fellowship to M.B.H. (W81XWH-13-1-0009) and NIH grants U54 CA163123, P01 HL024136 and R21CA167601.

**Author Contributions** M.B.H. designed and conducted most of the experiments, data analysis, and drafted the manuscript; A.B. developed the Intercostal Imaging Window; A.N. generated B16ZsGreen cell line and discussed data; E.W.R. generated mTmG mouse embryonic fibroblasts and provided editorial support on manuscript; M.L.R. participated in development and troubleshooting of LIVM method, discussed data and project direction; A.G. participated in intravital LN experiments and provided editorial support on manuscript. M.F.K. designed experiments, interpreted data, and with other authors, developed the completed manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.F.K. (matthew.krummel@ucsf.edu).

## METHODS

**Mice.** Mice were housed and bred under specific pathogen-free conditions at the University of California, San Francisco Laboratory Animal Research Center and all experiments conformed to ethical principles and guidelines approved by the UCSF Institutional Animal Care and Use Committee. C57/BL6 mice were purchased from Simonsen Laboratories or bred in house, and unless otherwise noted animals used were male between 6–8 weeks of age. Actin–CFP[20] mice were obtained from I. Weissman (Stanford University). MacBlue[14] mice were a gift from D. Hume (The Roslin Institute). CD11c-mCherry[21] mice were a gift from L. Lefrancois (University of Connecticut). CCR2-knockout[22] mice were a gift from J. Cyster. mTmG[23], Nur77-GFP[24], and Zbtb46-DTR[18] mice were purchased from Jackson Laboratories.

**Cell lines.** B16F10 (ATCC), PyMT-B[25,26] was a gift from J. Massegue and S. Abrams, MDA-MB231-expressing GFP was a gift from Z. Werb, mouse embryonic fibroblasts were prepared from mTmG mice as follows. Day 13.5 embryos were collected from pregnant mTmG females. Following removal of fetal liver embryos, were minced and subjected to Trypsin digestion. The retrieved cells were washed and resuspended in DMEM (GIBCO) and 10% heat-inactivated FCS and L-glutamate with penicillin and streptomycin. Cells were plated and passaged overnight. Media was aspirated after 24 h to remove any cells remaining in suspension and replaced with fresh media. Cells were then grown to 70%–80% confluency and cryopreserved. For experimental use, a vial of mouse embryonic fibroblasts were thawed and grown to 80% confluency in fresh media and used immediately for experiments. ZsGreen- and DsRed-expressing cells lines were generated by retroviral transduction with empty pSiren–ZsGreen (Clontech) or pSiren–DsRED(Clontech). Retrovirus was generated in Phoenix packaging cells (as previously described[27]) and applied to sub-confluent B16F10 or PyMT-B cells. Transduced cells were sorted for fluorescent-protein-positive cells on day 2 after infection on a FACSAria III sorter. Following an additional week of culture, cells were sorted a second time to ensure faithful expression of the reporter. Cell lines were subsequently tested and confirmed to be mycoplasma free by PCR.

**Metastasis induction.** For intravital imaging, flow cytometry and T cell activation assay sorting experiments, cultured B16ZsGreen cells were collected with Trypsin/EDTA and cultured for 30 min at 37 °C in complete media with or without Hoechst-3342 (Molecular Probes) at 1 μg ml$^{-1}$ to stain for nuclei and/or recover cells after collection. Cells were subsequently washed ×2 in PBS and $5 \times 10^5$ cells were injected via the tail vein into mice. For metastasis quantification experiments between $1 \times 10^5$ and $2.5 \times 10^5$ B16ZsGreen cells, prepared in the same fashion, were injected via the tail vein.

**Intravital imaging of pulmonary metastasis via intercostal insertion window.** This is a modified version of our previously published method of stabilized lung imaging[6], modifications are as follows. Mice were anaesthetized with 2.5% Avertin at a dose of 10 μl g$^{-1}$ and secured with tape to a custom heated microscope stage. Tracheostomy was performed to insert a small tracheal cannula, which was sutured into place and attached to a MiniVent mouse ventilator (Harvard Apparatus). Mice were ventilated with a stroke volume of 10 μl of compressed air (20–22% $O_2$) per gram of mouse weight, a respiratory rate of 130–140 breaths per minute, and a positive-end expiratory pressure of 2.5–3 cm $H_2O$. Isoflurane was continuously delivered at 1.5% to maintain anaesthesia and mice were given i.v. Lactated Ringers Solution (Baxter Health Care) at a rate of 0.8–1.6 μl min$^{-1}$ continuously during imaging. The mice were then placed in the right lateral decubitus position and a small surgical incision was made to expose the rib cage. A second incision was then made into the intercostal space between ribs 4 and 5, through the parietal pleura, to expose the surface of the left lung lobe. A flanged thoracic suction window with 8 mm coverslip (Extended Data Fig. 2) was then inserted between the two ribs and secured to the stage using a set of two optical posts and a 90° angle post clamp (Thor Labs). 20–25 mm Hg of suction was applied (Amvex Corporation) to gently immobilize the lung. The two-photon microscope objective was then lowered into place over the thoracic suction window. For imaging of the arrival of metastatic tumour cells, $5 \times 10^5$ B16F10 cells expressing ZsGreen were injected inline through the i.v. line during imaging. In experiments where metastatic cells were imaged at later time points (6–24 h), cells were instead injected i.v. through a tail vein injection at the appropriate time point before performing the intravital imaging surgery.

**Two-photon microscopy.** Intravital imaging was performed using a custom-built two-photon setup equipped with two infrared lasers (MaiTai, Spectra Physics; Chameleon, Coherent). The MaiTai laser was tuned to 810 nm for excitation of CFP. Chameleon laser excitation was tuned to 980 nm for simultaneous excitation of TdTomato or mCherry and ZsGreen and/or GFP. Emitted light was detected using a 25× 1.2NA water lens (Zeiss) coupled to a 6-colour detector array (custom; using Hamamatsu H9433MOD detectors). Emission filters used were: blue 475/23, green 510/42, yellow 542/27, red 607/70, far red 675/67. The microscope was controlled by the MicroManager software suite, $z$-stack images were acquired with fourfold averaging and $z$-depths of 3 μm Data analysis was performed using the Imaris software suite (Bitplane).

**Non-shear lung slice imaging.** Slice imaging was performed as previously described[13]. In brief, mice were injected with $5 \times 10^5$ B16ZsGreen cells by tail vein. After 1 h mice were euthanized by anaesthetic overdose (1 ml of 2.5% Avertin), mice were then intubated by tracheotomy with the sheath from an 18-gauge i.v. catheter. Lungs were subsequently inflated with 1 ml of 2% low melting temp agarose (BMA) in sterile PBS at 37 °C. Agarose was then solidified by flooding the chest cavity with 4 °C PBS. Inflated lungs were excised and the left lobe was cut into ∼300-μm sections using a vibratome. Sections were mounted on plastic coverslips and imaged by two-photon microscopy at 37 °C in carbogen (5% $CO_2$:95% $O_2$)-perfused RPMI-1640 media (Gibco, without Phenol Red).

**Lymphnode explant imaging.** Explant imaging was performed as previously described[28]. In brief, 2 million GFP-expressing OT-I T cells were transferred in CD11c-mCherry mice 1 day before injection with $5 \times 10^5$ B16ZsGreenSL8 cells via tail vein. Mediastinal lymph node was removed, cleaned of fat, and immobilized on a plastic coverslip with the hilum facing away from the objective. Lymph nodes were imaged in 30-min intervals with 810 nm excitation on a two-photon microscope, as above.

**Tissue digests for flow cytometry and sorting.** Lungs were collected from mice following euthanasia by overdose with 2.5% Avertin. Lungs were placed in 5 ml of DMEM (GIBCO) with 0.26 U ml$^{-1}$ LiberaseTM (Roche) and 0.25 mg ml$^{-1}$ DNaseI (Roche). Samples were placed in C-Tubes (Miltenyi) and briefly processed with a GentleMACS Dissociator (Miltenyi). Samples were then incubated at 37 °C for 30 min and processed a second time via GentleMACS. Tissue homogenate was then passed through a 100 μm Nytex Filter. Red blood cells were lysed with 3 ml of 175 mM $NH_4Cl$ per lung for 5 min at 37 °C. Samples were then filtered through a 40 μm Nytex filter and resuspended for subsequent FACS staining. For experiments where vascular localization was assessed, mice were injected with i.v. 10 μg CD45–APC (allophycocyanin; eBioscience Clone 30-F11) 5 min before collection and prepared as per the previously published protocol[15]. Lymph nodes were prepared by placing in 1 ml of digestion buffer and puncturing with sharp forceps. Samples were incubated for 15 min at 37 °C and then pipetted up and down ∼20 times with a P1000 pipette to dissociate. Samples were returned to 37 °C for 15 min. Cell suspension was then filtered through a 40 μm Nytex filter and prepared for subsequent FACS staining.

**Flow cytometry.** For surface staining, cells were incubated with anti-Fc receptor antibody (clone 2.4G2) and stained with antibodies in PBS and 2% fetal calf serum for 30 min on ice. Viability was assessed by staining with fixable Live/Dead Zombie NIR (Biolegend) or 4′,6-diamidino-2-phenylindole (Molecular Probes). All flow cytometry was performed on a BD Fortessa flow cytometer. Analysis of flow cytometry data was performed using Flowjo (Treestar). Cell sorting was performed using a BD FACS Aria II or a BD FACS Aria III.

Antibody clones used in these studies: CD45-Alexa 700 (eBioscience clone 30-F11), MHCII Brilliant Violet 421 (Biolegend clone M5/114.15.2), CD11c Brilliant Violet 510 (Biolegend clone N418), CD11b Brilliant Violet 605 or PerCPCy5.5 (Biolegend clone M1/70), Ly6C Brilliant Violet 711 (Biolegend clone HK1.4), CD24 PeCy7 (Biolegend clone M1/69), CD103 PE (Biolegend clone 2E7), CD69 PeCy7 (H1.2F3) Ly6G A647 (Biolegend 1A8), Siglec-F A647 (BD Bioscience clone E50-2440), CD19 A647 (Biolegend clone 6D5), CD90.2 A647 (Biolegend clone 30.H12), VCAM-1 PE (Biolegend clone 429), CD88 PE (Biolegend clone 20/70), CD155 PE (Biolegend clone 4.24.1), CMKLR1 PE (Biolegend clone BZ194), CD38 PE (Biolegend clone 14.27), CD63 PE (Biolegend clone NVG-2) was performed in FACS buffer for 20 min at 4 °C. Live/Dead discrimination was performed using either Propidium Iodide or Zombie NIR Fixable Live/Dead Stain (Biolegend). Samples were analysed on a LSRFortessa (BD Biosciences) in the UCSF Flow Cytometry Core.

**T-cell activation assay.** Assays were performed as previously published[19]. In brief, lymph node cells from OT-I TCR transgenic mice were isolated and were enriched for naive CD8$^+$ T cells by StemSep CD8 enrichment (Stemcell Technologies). Naive OT-I CD8 cells or labelled with 2 μM succinimidyl ester eFluor 670 (SE670, eBioscience) and mixed with APC at a 10:1 ratio (1,0000 T cells: 1,000 APC) with unpulsed mLN antigen-presenting cells (sorted from the lymph node 72 h following injection with either B16ZsGreenSL8 or B78mCherryOVA) in 96-well V-bottom plates for either 24 or 72 h at 37 °C in 5% $CO_2$, at which point activation was measured by CD69 and Nur77-GFP upregulation and SE670 dilution using flow cytometry.

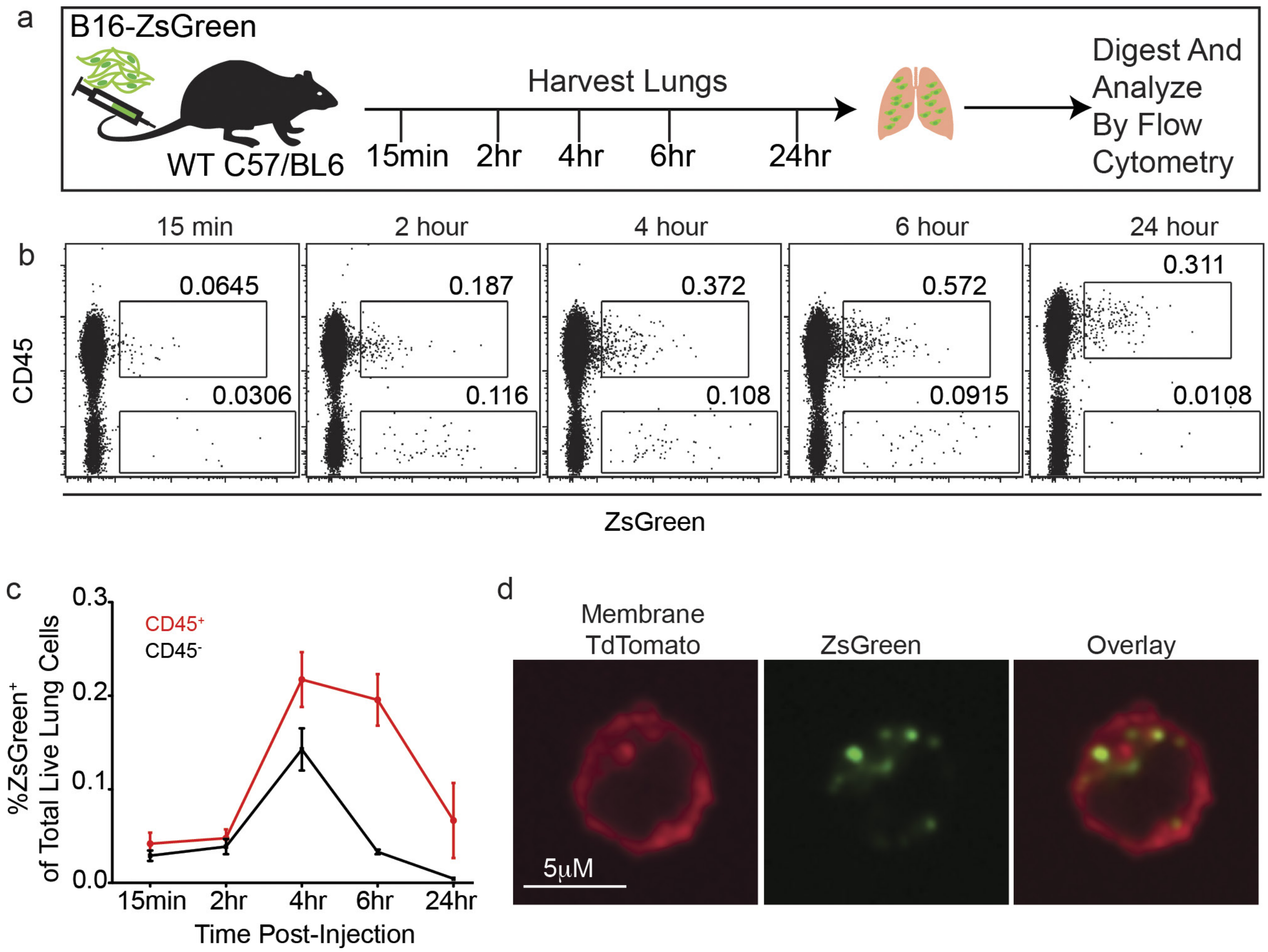
**Zbtb46-DTR bone marrow chimaeras.** In order to avoid mortality associated with diphtheria-toxin-mediated depletion of Zbtb46-DTR expressing cells in intact animals[18], we generated bone marrow chimaeras. Bone marrow cells ($3 \times 10^6$) from Zbtb46-DTR mice were adoptively transferred retro-orbitally into lethally irradiated recipients. Animals were allowed to recover and repopulate the haematopoietic compartment for 8 weeks. On day 1, animals were injected with

20 ng of diphtheria toxin in PBS or an equivalent volume of PBS alone. On day 0, animals were injected with 150,000 B16ZsGreen cells via tail vein injection. On day +1, animals were injected with a second dose of 20 ng of diphtheria toxin or PBS. Metastases were then allowed to develop for 2 weeks before quantification, as described elsewhere in these methods. In alternate experiments, mice were first injected with $1 \times 10^5$ B16F10 cells in growth factor reduced Matrigel (Corning, 50 μl of 50% Matrigel diluted in PBS) 3 weeks before diphtheria toxin and metastasis injection injection as per the above protocol. Primary tumours were not allowed to exceed 2 cm before euthanasia in accordance with University of California, San Francisco Institutional Animal Care and Use Committee protocol, as such in these experiments metastases were collected at 1 week to fit within bounds of primary tumour endpoint criteria.

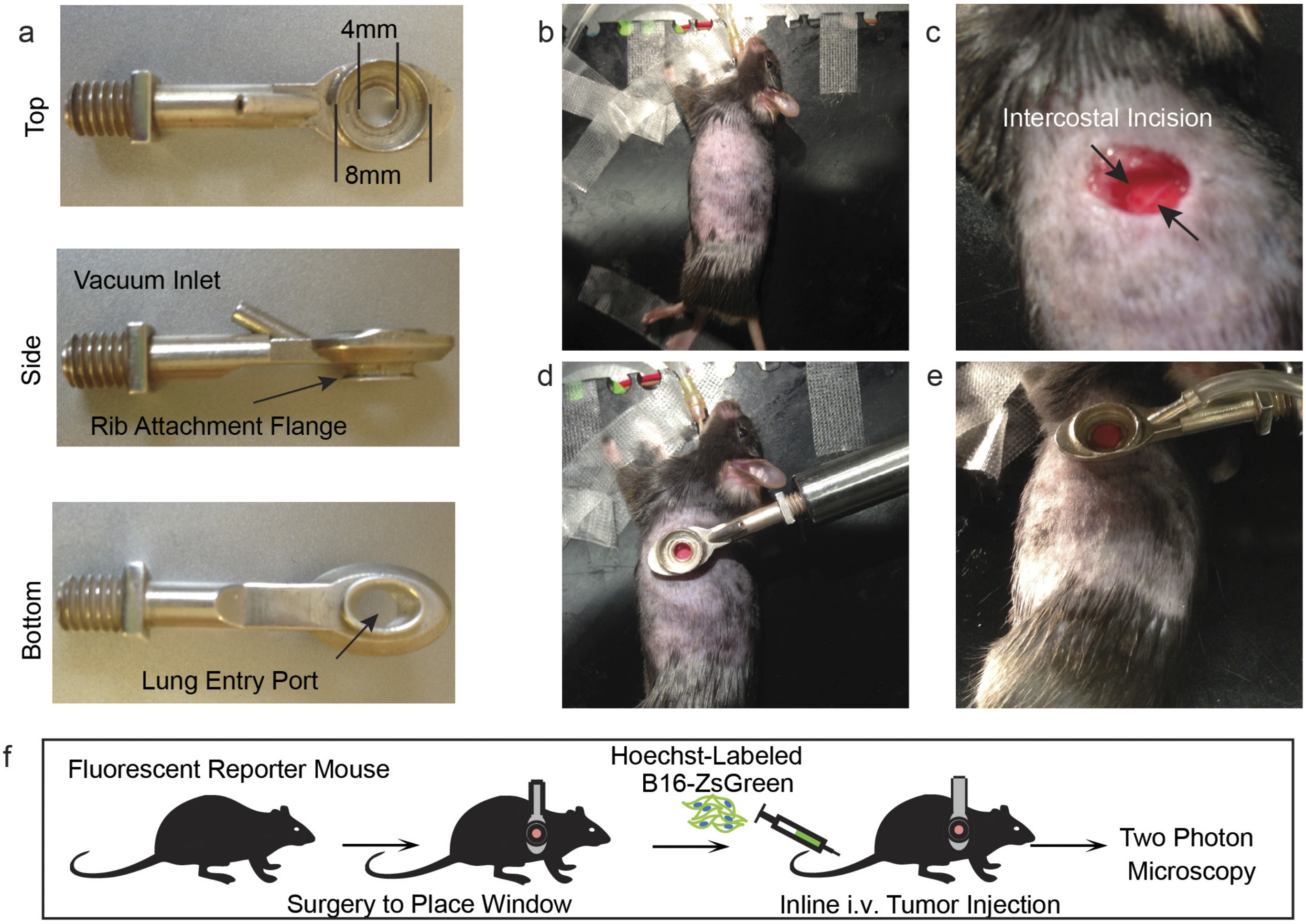
**Assessment of cell activity.** Data for tumour cells of interest collected from intravital imaging of lung was used for these analyses. Tumour cells were assessed for a 30 min interval at either 2 h or 24 h after injection. Data were first subjected to a maximum intensity projection and output data was binarized at each time point to reveal the maximum boundaries of the cell at any given time point. Data was then projected over time to yield an image depicting the total area occupied by the cell over the 30 min window. Areas were then calculated for the total area of the time projection, as well as the common area colocalized between the initial time point (T0), the final time point (T30) and the time projection (T-Project) to yield the parameter T-Overlap. The cell activity index was then calculated as Cell Activity Index = T-Project/T-Overlap.

**Image and statistical analysis.** All image analysis was performed using Imaris (Bitplane) in conjuction with Matlab (Mathworks). For statistical analysis, unless otherwise noted, all data were expressed as mean ± s.d. Comparisons between groups were analysed with by *t*-test, multi-group comparisons were performed using a one-way ANOVA test and Bonferroni post-hoc test, using Graphpad Prism software. In all cases where statistical significance is provided, variance was not statistically different between groups with the exception of Fig. 1h. Sample sizes were chosen on the basis of previous experience in the lab with respect to inherent variability in i.v. metastatic models and intravital imaging. Animals within each cohort were randomly assigned to treatment groups. In cases where wild-type mice (Simonsen) were used, entire cages were randomly assigned into treatment groups. Blinded analysis was not performed in these studies. Data were considered as statistically significant when $P$ values were $<0.05$, indicated by '*' in figures, unless otherwise noted. $t_{1/2}$ value was calculated as the half-life (ln(2)/(K); where $K$ is the rate constant of the regression) of a one-phase exponential decay nonlinear regression.

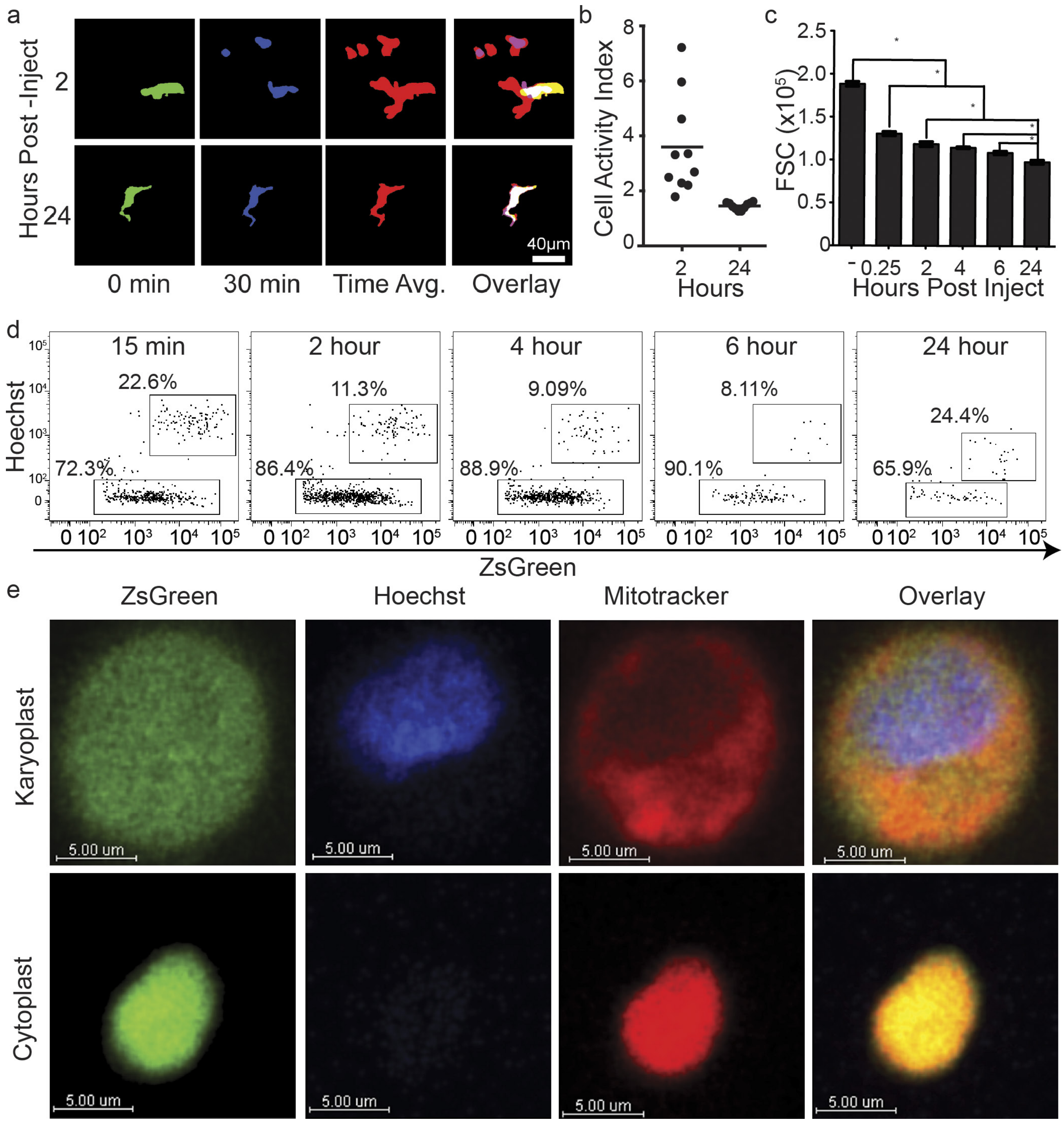
20. Hadjantonakis, A. K., Macmaster, S. & Nagy, A. Embryonic stem cells and mice expressing different GFP variants for multiple non-invasive reporter usage within a single animal. *BMC Biotechnol.* **2,** 11 (2002).
21. Khanna, K. M. *et al.* T cell and APC dynamics in situ control the outcome of vaccination. *J. Immunol.* **185,** 239–252 (2010).
22. Boring, L. *et al.* Impaired monocyte migration and reduced type 1 (Th1) cytokine responses in C–C chemokine receptor 2 knockout mice. *J. Clin. Invest.* **100,** 2552–2561 (1997).
23. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45,** 593–605 (2007).
24. Moran, A. E. *et al.* T cell receptor signal strength in $T_{reg}$ and iNKT cell development demonstrated by a novel fluorescent reporter mouse. *J. Exp. Med.* **208,** 1279–1289 (2011).
25. Stewart, T. J. & Abrams, S. I. Altered immune function during long-term host-tumor interactions can be modulated to retard autochthonous neoplastic growth. *J. Immunol.* **179,** 2851–2859 (2007).
26. Acharyya, S. *et al.* A CXCL1 paracrine network links cancer chemoresistance and metastasis. *Cell* **150,** 165–178 (2012).
27. Friedman, R. S., Jacobelli, J. & Krummel, M. F. Surface-bound chemokines capture and prime T cells for synapse formation. *Nature Immunol.* **7,** 1101–1108 (2006).
28. Gérard, A. *et al.* Secondary T cell–T cell synaptic interactions drive the differentiation of protective $CD8^+$ T cells. *Nature Immunol.* **14,** 356–363 (2013).



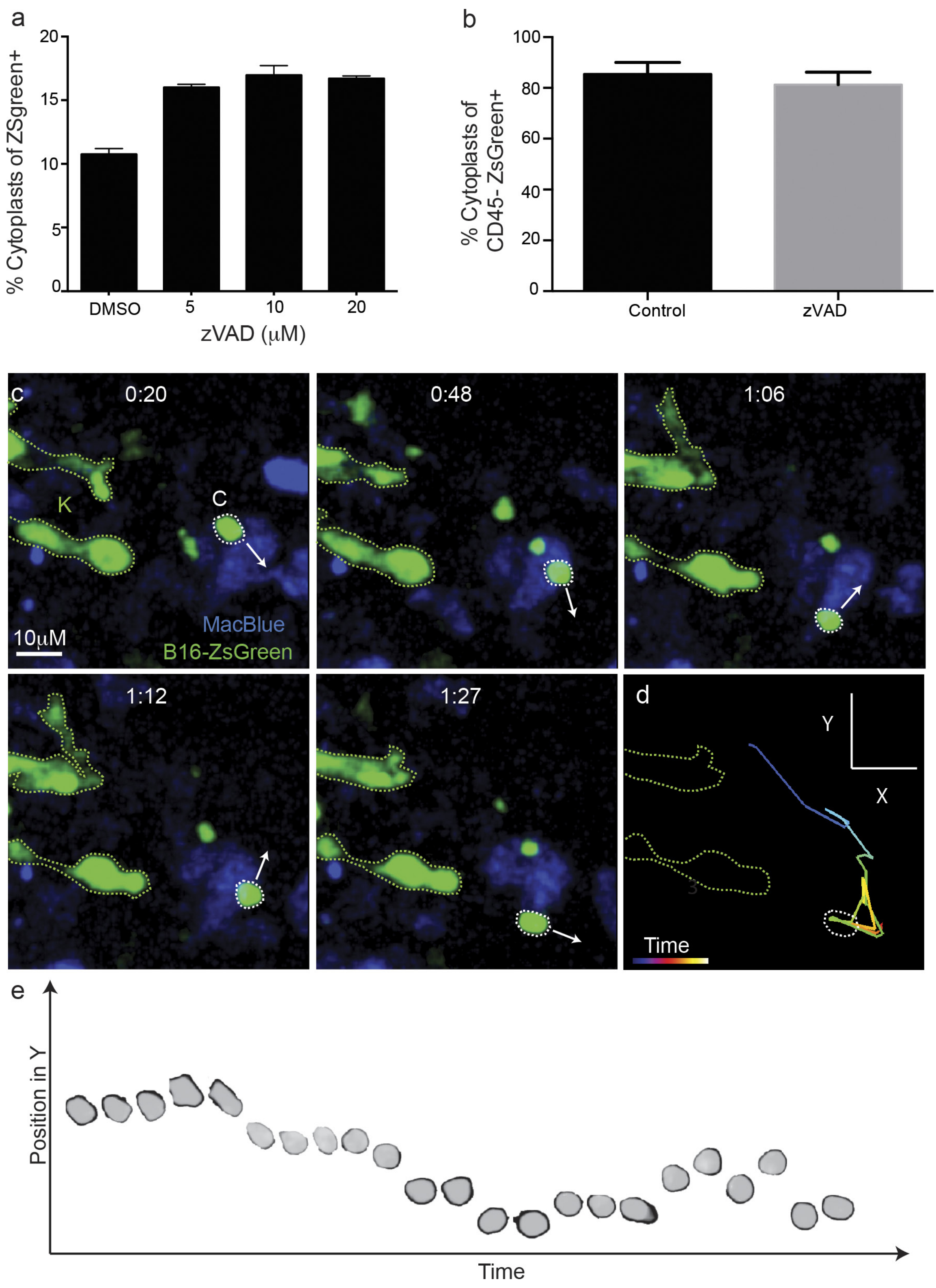
**Extended Data Figure 1 | Loading of lung immune cells by prospective metastatic B16 melanoma. a**, Schema for assessing immune cell loading by i.v.-injected B16ZsGreen cells. **b**, Representative plots of ZsGreen$^+$ populations in the lungs of mice injected with B16ZsGreen over the first 24 h following injection. Data are gated on the basis of expression of the immune cell marker CD45. **c**, Quantification of **b** with $n = 6$ per group, error bars are s.d. **d**, Confocal imaging of CD45$^+$ZsGreen$^+$ cells sorted from a lung digest from a ubiquitous membrane-bound TdTomato fluorescent mouse 24 h after i.v. injection with B16ZsGreen.



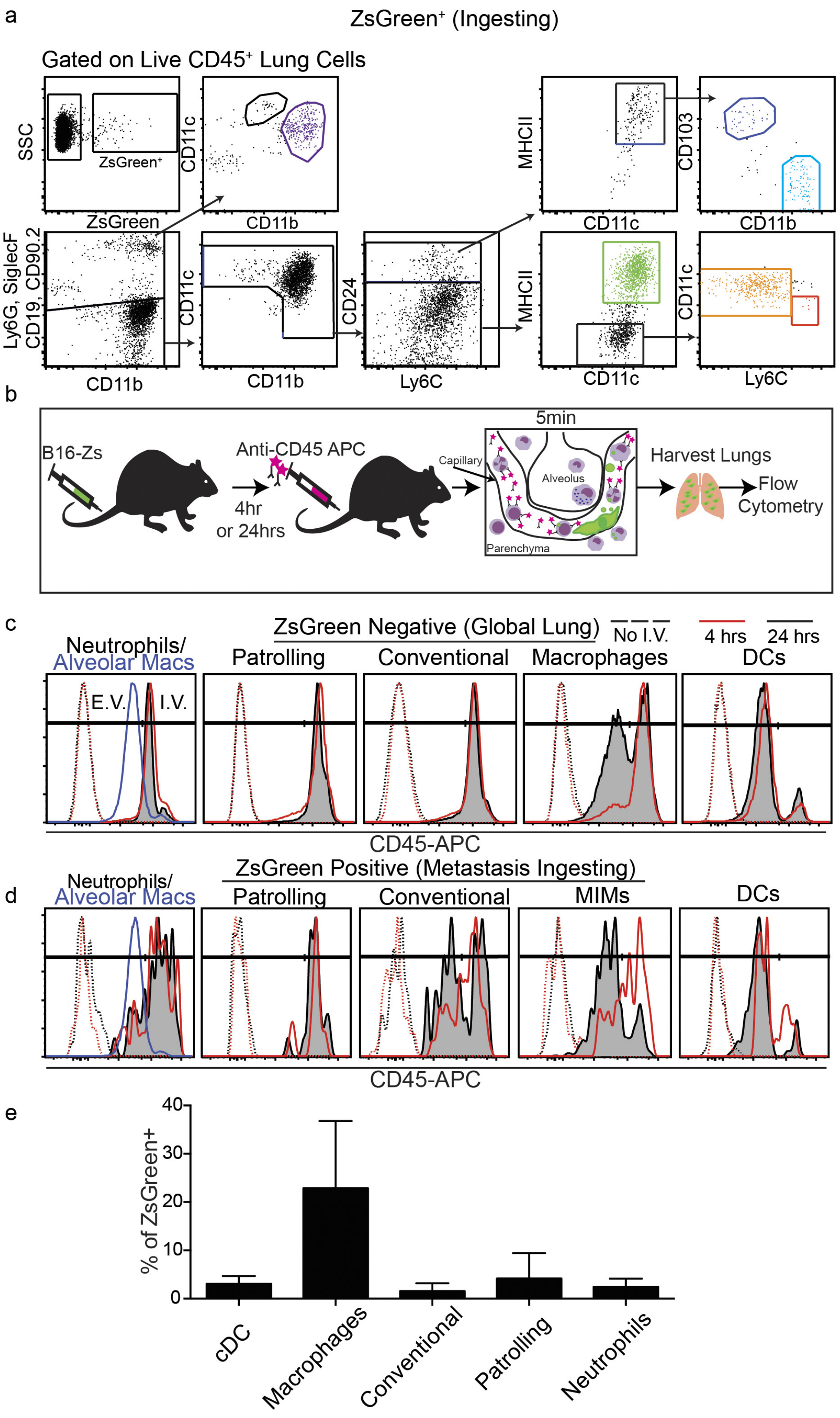
**Extended Data Figure 2 | Intercostal insertion window for lung intravital microscopy.** **a**, Top, side, and bottom views of the intercostal insertion window. The window accommodates an 8 mm coverslip and allows for visualization of a 4 mm field of the left lung lobe. **b**–**e**, Images detailing surgical insertion of the intercostal window. **b**, Mouse is intubated, attached to ventilator, and placed in right lateral decubitis position and surgical field is shaved. **c**, An ~6 mm incision is made immediately above ribs 4 and 5 over the anterior surface of left lung lobe. **d**, The intercostal window is slipped between ribs 4 and 5, and attached to a rigid support. **e**, Around 20 mm Hg of vacuum suction is applied to the window to secure a small portion of lung to the coverglass. **f**, Schema showing approach for two-photon intravital microscopy of lung seeding by B16ZsGreen cells.



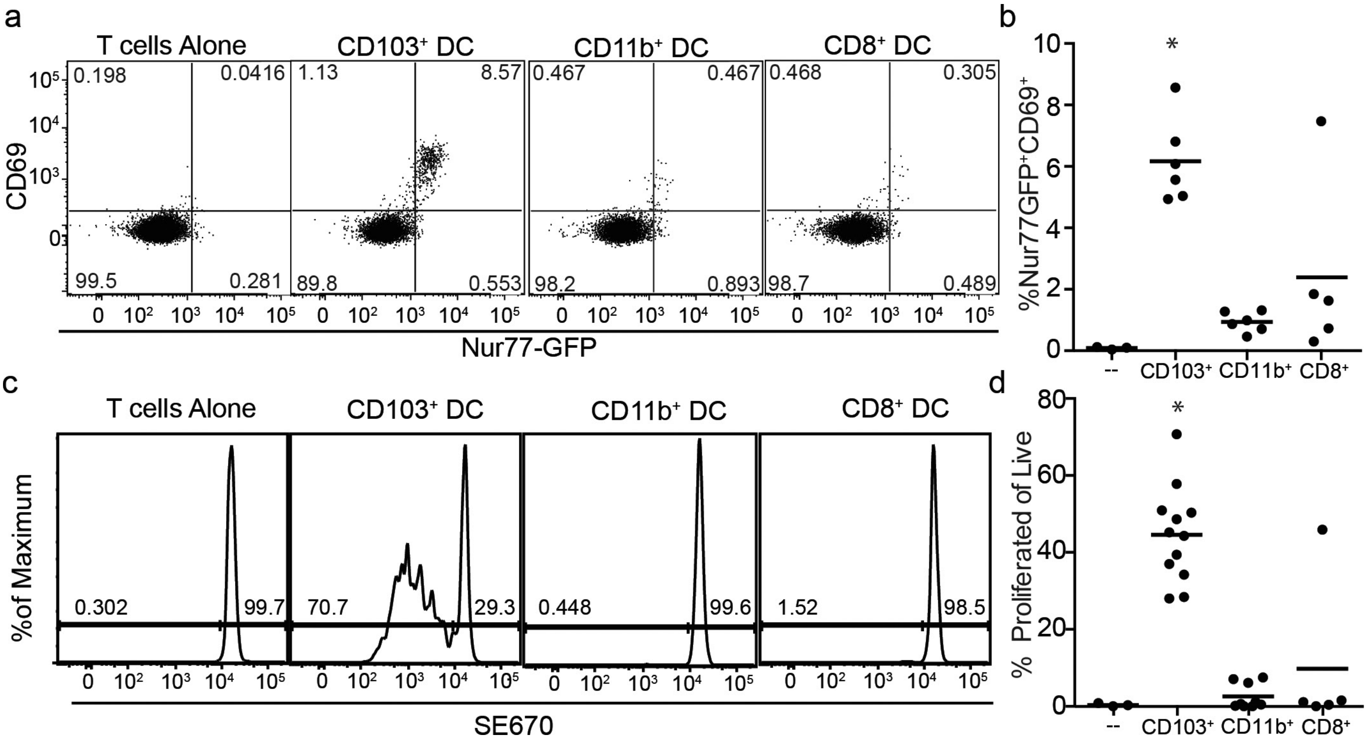
**Extended Data Figure 3 | Assessment of tumour cell activity and cytoplast characterization. a**, Binarized maximum intensity projections of representative cells at 2 or 24 h after injection. Cells were time-projected over a 30 min window to assess overall cellular activity during the interval. Images show the beginning time point (0 min), the ending time point (30 min), the time projection, and the overlay. White-filled space in the overlay represents the region of the cell that was stable during the analysed interval. These data are associated with Supplementary Video 3. These data are representative from imaging performed in 3 mice. **b**, Quantification of the cell activity index (defined as the area calculated from the projection of all positions over time as a ratio of the common area over all time (for example, area of overlap)) from analysis in **d** ($n = 10$ cells per group, horizontal bars are mean value). **c**, Flow cytometric quantification of FSC for tumour cells isolated from lung via digestion at 15 min, 2 h, 4 h, 6 h, and 24 h after injection. $*P < 0.05$, one-way ANOVA with Bonferroni post-hoc test, error bars are s.d. **d**, Representative data from flow cytometric discrimination of nucleated karyoplasts and anucleate cytoplasts derived from B16ZsGreen tumour cells in the lung *in vivo* over a 24 h time course, full data set quantified in Fig. 1e. **e**, confocal analysis of Hoechst and Mitotracker-labelled B16ZsGreen karyoplasts and cytoplasts sorted from *in vitro* culture of B16ZsGreen cells.



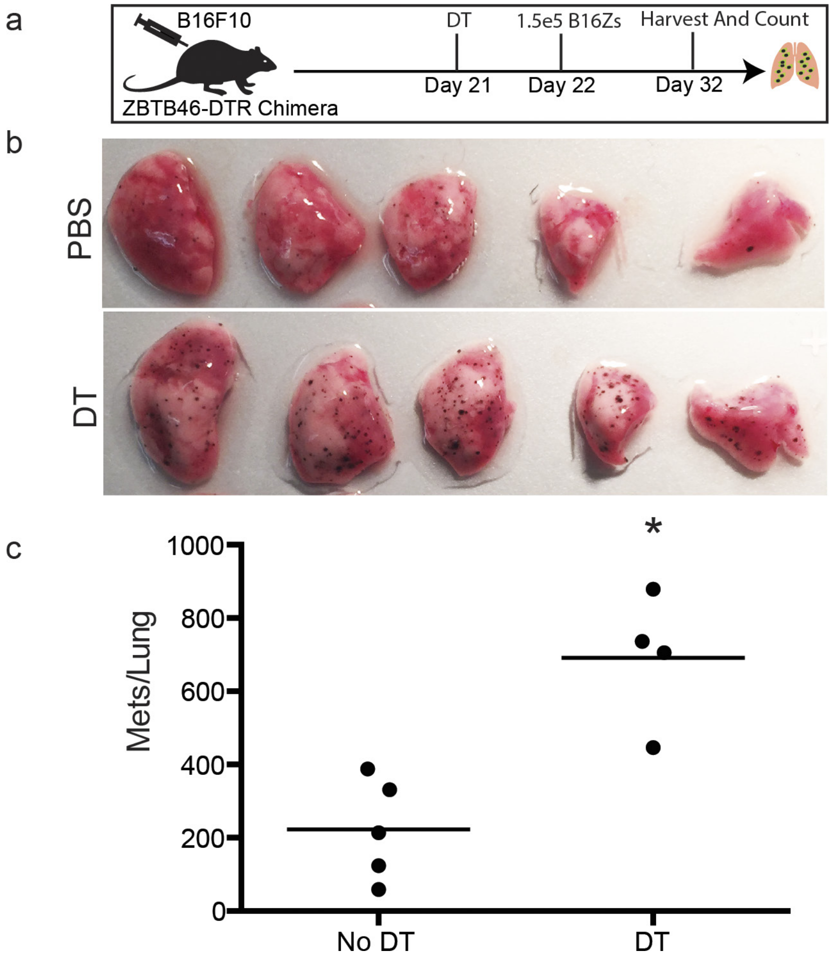
**Extended Data Figure 4 | Autonomous motility of cytoplasts *in vivo*.** **a**, Flow cytometric quantification of cytoplasts *in vitro* following 24 h treatment with Z-VAD at indicated concentrations. **b**, Flow cytometric quantification of cytoplasts *in vivo* from lung digests of mice treated with 10 μg Z-VAD i.v at the time of injection with 2.5 × $10^5$ B15ZsGreen cells (**a** and **b**, $n = 4$ per group; no significant difference detected between groups, unpaired *t*-test, error bars are s.d.). **c**, Image series for B16ZsGreen cytoplast migrating autonomously through the lung microvasculature of a MacBlue host. Arrows represent the direction of the trajectory of the cytoplast at indicated time point. These data are associated with Supplementary Video 6. These data are representative of imaging collected from at least 12 mice. **d**, Representative tracking of a cytoplast from Supplementary Video 6 (and Extended Data Fig. 4c). **e**, A superposition image of 23 consecutive time points of a cytoplast migrating through lung microvasculature. Image shows the change in position in the *y* axis of direction as defined in **d** at each subsequent timepoint as the cytoplast migrates up and down the vessel.



**Extended Data Figure 5 | Characterization of tumour-interacting myeloid cell waves. a**, Representative gating for total lung myeloid populations. **b**, Schema detailing method for discrimination of intravascular versus extravascular localization of lung myeloid populations. **c, d**, Representative histograms of intravascular CD45 staining used to discriminate between intravascular and extravascular localization of lung myeloid cells at 4 and 24 h after injection with B16ZsGreen. Data quantified in Fig. 3d. In the leftmost panels, alveolar macrophages at 24 h after tumour injection are shown as a known control for extravascular staining. **c**, Total lung myeloids. **d**, ZsGreen$^+$ myeloid cells. **e**, Quantification ZsGreen$^+$ myeloid populations by flow cytometry in lungs of mice bearing 2-week subdermal B16ZsGreen tumours ($n = 4$ per group). Error bars are s.d.



**Extended Data Figure 6 | Stimulatory capacity of CD103$^+$ DCs in lung-draining lymph node. a**, CD69 vs Nur77-GFP expression 24 h after culture from *ex vivo* coculture of OT-I TCR transgenic CD8$^+$ T cells with sorted APCs from mLNs, where the latter were isolated 72 h post-injection with B16ZsGreenSL8. **b**, Quantification of **a**. **c**, Dilution of SE670 as an index of proliferation 72 h after culture with indicated APC populations. **d**, Quantification of **c**. $n = 6$ (**b**) or 6–12 (**d**) per group from 2 experiments; $*P < 0.05$, one-way ANOVA with Bonferroni post-hoc test, horizontal bars are mean value.



**Extended Data Figure 7 | cDCs confer anti-metastatic activity in the presence of a primary tumour. a**, Experimental schema for evaluation of role of cDCs in lung metastasis in the presence of a primary tumour. **b**, Representative images of lungs from Zbtb46-DTR bone marrow chimaeras treated with PBS or DT after implantation of a primary subdermal tumour ($1 \times 10^5$ B16F10 in matrigel) and i.v. metastases ($1.5 \times 10^5$ B16ZsGreen). Metastases were assessed one week after i.v. injection. **c**, Quantification of total number of visible ZsGreen$^+$ lung metastases in PBS- or DT-treated Zbtb46-DTR bone marrow chimaeras. $n = 4$–5 per group, representative of 2 experiments; $*P < 0.05$, unpaired *t*-test, horizontal bars are mean value.

doi:10.1038/nature17161

# Melanoma addiction to the long non-coding RNA *SAMMSON*

Eleonora Leucci[1,2], Roberto Vendramin[1,2], Marco Spinazzi[2], Patrick Laurette[3], Mark Fiers[2], Jasper Wouters[4], Enrico Radaelli[5], Sven Eyckerman[6,7], Carina Leonelli[8,9], Katrien Vanderheyden[8,9], Aljosja Rogiers[1,2], Els Hermans[10], Pieter Baatsen[2], Stein Aerts[11], Frederic Amant[10], Stefan Van Aelst[12,13], Joost van den Oord[4], Bart de Strooper[2], Irwin Davidson[3], Denis L. J. Lafontaine[14], Kris Gevaert[6,7], Jo Vandesompele[8,9], Pieter Mestdagh[8,9]* & Jean-Christophe Marine[1,2]*

**Focal amplifications of chromosome 3p13–3p14 occur in about 10% of melanomas and are associated with a poor prognosis. The melanoma-specific oncogene *MITF* resides at the epicentre of this amplicon[1]. However, whether other loci present in this amplicon also contribute to melanomagenesis is unknown. Here we show that the recently annotated long non-coding RNA (lncRNA) gene *SAMMSON* is consistently co-gained with *MITF*. In addition, *SAMMSON* is a target of the lineage-specific transcription factor SOX10 and its expression is detectable in more than 90% of human melanomas. Whereas exogenous *SAMMSON* increases the clonogenic potential in *trans*, *SAMMSON* knockdown drastically decreases the viability of melanoma cells irrespective of their transcriptional cell state and *BRAF*, *NRAS* or *TP53* mutational status. Moreover, *SAMMSON* targeting sensitizes melanoma to MAPK-targeting therapeutics both *in vitro* and in patient-derived xenograft models. Mechanistically, *SAMMSON* interacts with p32, a master regulator of mitochondrial homeostasis and metabolism, to increase its mitochondrial targeting and pro-oncogenic function. Our results indicate that silencing of the lineage addiction oncogene *SAMMSON* disrupts vital mitochondrial functions in a cancer-cell-specific manner; this silencing is therefore expected to deliver highly effective and tissue-restricted anti-melanoma therapeutic responses.**

*In silico* analysis of single nucleotide polymorphism (SNP) array data from >300 human clinical samples from The Cancer Genome Atlas (TCGA) revealed that the chromosome 3p melanoma-specific focal amplifications invariably encompass a recently annotated long intergenic non-coding RNA (lincRNA) gene, *SAMMSON*, which is located $\sim$30 kilobase pair (kb) downstream of *MITF* (Fig. 1a and Extended Data Fig. 1a, b). Importantly, *SAMMSON* expression levels correlated with copy number gain ($P < 0.001$; Fig. 1a). Genome-wide copy number analysis (CNA) of melanoma cell lines and short-term cultures (MM lines[2]) confirmed co-amplification of *MITF* and *SAMMSON* in a subset of these samples (Extended Data Fig. 1c).

Unexpectedly, although *MITF-SAMMSON* co-amplification only occurs in about 10% of melanomas, analysis of the TCGA RNA-sequencing (RNA-seq) data set detected *SAMMSON* in >90% of both primary and metastatic skin cutaneous melanomas (SKCMs; Fig. 1b and Extended Data Fig. 1a). *SAMMSON* was also detected in 16 out of 17 randomly selected MM lines (Extended Data Fig. 1d). MM001 was the only line that did not express *SAMMSON* owing to a decrease in copy number (Extended Data Fig. 1c, d). *SAMMSON* levels were comparable in cultures that exhibited a 'proliferative' and 'invasive' transcriptional profile (Extended Data Fig. 1d)[3]. In contrast, MITF-M protein levels were high in the proliferative and low in the invasive cultures, respectively. Thus, the levels of *SAMMSON* and MITF do not strictly correlate. Likewise, there was no correlation between *SAMMSON* and *MITF* expression in the melanoma TCGA clinical samples (Extended Data Fig. 1e, f). Furthermore, there was no correlation between *SAMMSON* expression and any of the common melanoma somatic mutations such as *BRAF*, *NRAS* or *TP53* (data not shown).

While barely detectable, if at all, in normal human melanocytes (NHMEs) and in non-invasive melanoma lesions in the radial growth phase (RGP), *SAMMSON* expression was readily detectable in invasive vertical growth phase (VGP) samples (Fig. 1c). This indicates that *SAMMSON* expression is specifically induced as melanoma cells transit from an immortalized to a fully transformed cell state and is therefore a putative biomarker of melanoma malignancy.

Quantitative polymerase chain reaction with reverse transcription (RT–qPCR) analysis in 60 different cancer cell lines detected *SAMMSON* exclusively in the melanoma samples (Extended Data Fig. 1g, h). Analysis of RNA-seq data from a total of 8,085 tumour specimens from 24 cancer types (TCGA) confirmed selective expression in melanoma (Fig. 1d). Whereas *SAMMSON* expression was detectable in normal human melanoblasts, it was undetectable in NHME cultures and in normal adult tissues (Fig. 1e and data not shown).

Consistent with *SAMMSON* being a lincRNA, the expression of which is driven by its own promoter, a peak of H3K4me3 chromatin immunoprecipitation followed by sequencing (ChIP-seq) was detected upstream of its transcription start site (TSS) in melanoma cells (Fig. 1f). Similarly, a peak of H3K27ac, a marker of active enhancers and promoters, overlaps with the H3K4me3 peak at the *SAMMSON* promoter. H3K27ac peaks are also found upstream of *SAMMSON* in all but one (MM001) MM lines (Extended Data Fig. 2a). Consistent with the melanoma-specific expression of *SAMMSON*, such H3K27ac peaks could not be detected in the non-melanoma cancer cell lines profiled by ENCODE.

Interestingly, we identified putative SOX-binding sites upstream of the *SAMMSON* TSS. SOX10 is a melanoblast/melanoma-specific transcription factor[4] and may therefore contribute to the melanoblast/melanoma-specific expression of *SAMMSON*. Consistently, SOX10 and its co-factor BRG1 (ref. 5), but not MITF, were recruited upstream of the *SAMMSON* TSS in melanoma cells (Fig. 1f, g and Extended Data Fig. 2b, c). Knockdown of SOX10, but not MITF, led to a decrease in *SAMMSON* expression (Fig. 1h and Extended Data Fig. 2d, e). These

[1] Laboratory For Molecular Cancer Biology, Center for Human Genetics, KULeuven, Herestraat 49, 3000 Leuven, Belgium. [2] Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [3] Institut de Génétique et de Biologie Moleculaire et Cellulaire (IGBMC), Rue Laurent Fries 1, 67404 Illkirch, France. [4] Laboratory of Translational Cell and Tissue Research, Department of Pathology, KULeuven and UZ Leuven, Herestraat 49, 3000 Leuven, Belgium. [5] Mouse Histopathology Core Facility, Center for the Biology of Disease, VIB-KULeuven, Herestraat 49, 3000 Leuven, Belgium. [6] Medical Biotechnology Center, VIB, Albert Baertsoenkaai 3, 9000 Gent, Belgium. [7] Department of Biochemistry, Gent University, Albert Baertsoenkaai 3, 9000 Gent, Belgium. [8] Center for Medical Genetics, Gent University, De Pintelaan 185, 9000 Gent, Belgium. [9] Cancer Research Institute Gent, Gent University, De Pintelaan 185, 9000 Gent, Belgium. [10] Gynaecologische Oncologie, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [11] Laboratory of Computational Biology, Center for Human Genetics, KULeuven, Herestraat 49, 3000 Leuven, Belgium. [12] Department of Applied Mathematics, Computer Science and Statistics, Gent University, De Pintelaan 185, 9000 Gent, Belgium. [13] Department of Mathematics, KU Leuven, Celestijnenlann 200B, 3001 Leuven, Belgium. [14] RNA Molecular Biology, Center for Microscopy and Molecular Imaging, Université Libre de Bruxelles (ULB), rue des Professeurs Jeener et Brachet 12, 6041 Charleroi, Belgium.
*These authors contributed equally to this work.

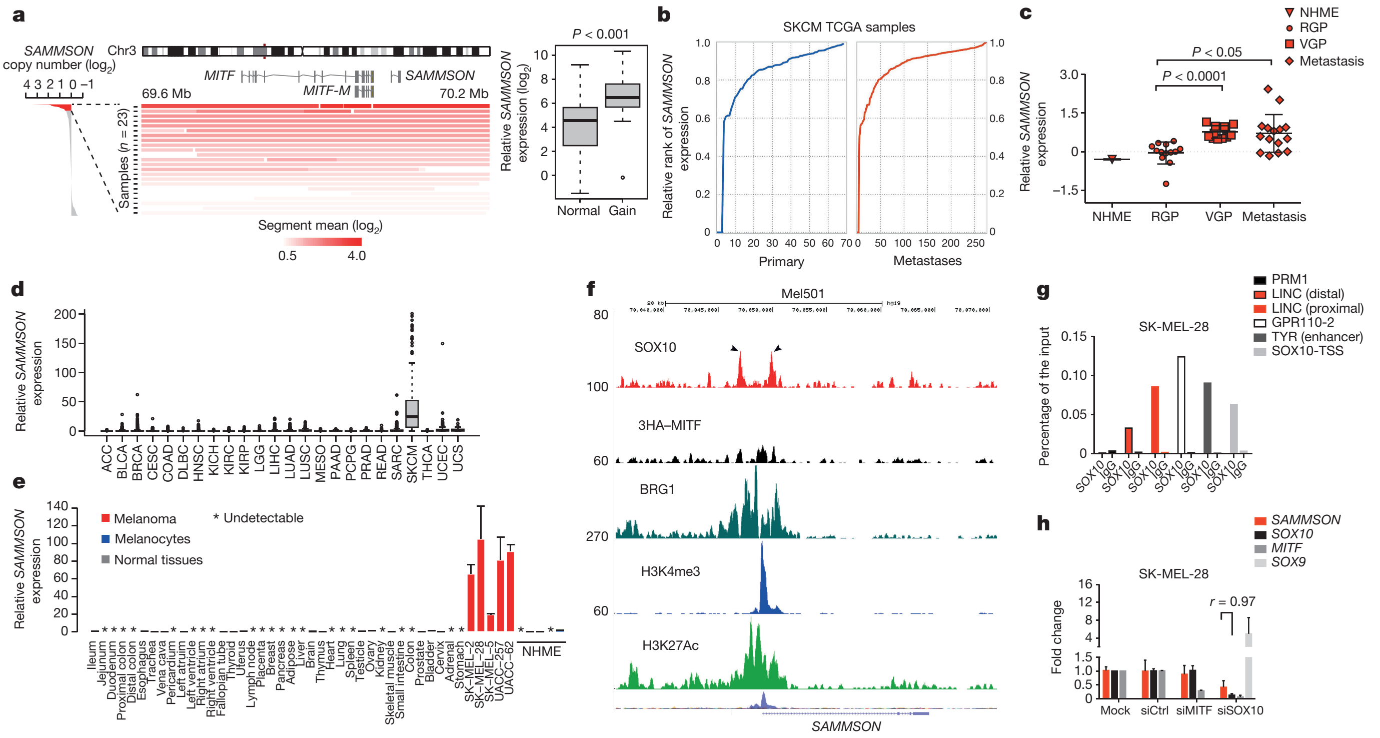


**Figure 1 | Gene amplification and SOX10-mediated transcription drives *SAMMSON* expression in melanoma.** **a**, DNA copy number of *SAMMSON* and *MITF* in melanoma lesions ($n = 386$). Samples are ranked according to *SAMMSON* copy number, expressed as the mean log ratio of the segment encompassing *SAMMSON*. Samples with a *SAMMSON* log ratio > 0.5 are shown (Mann–Whitney $P < 0.001$). Chr, chromosome. **b**, Read counts were generated based on the GENECODE annotation (version 19) from the RNA-seq alignments of 345 SKCM samples (TCGA). 7,014 lncRNAs were selected for which the raw counts were normalized to reads per kb per million reads (RPKM) values. The rank expression of *SAMMSON* was determined, sorted and plotted for the primary (left) and metastatic (right) samples. **c**, Relative expression in NHME, RGP, VGP and metastatic melanoma. Significance was calculated by analysis of variance (ANOVA). **d**, Relative expression across a cohort of 8,085 primary tumour samples from different cancer types. **e**, Relative expression across 38 adult tissues, 5 melanoma cell lines and NHME cultures. Error bars represent standard deviations (s.d.) of qPCR replicates ($n = 2$). **f**, University of California, Santa Cruz (UCSC) Genome Browser screenshots of ChIP-seq data illustrating the selective recruitment of SOX10 to the *SAMMSON* loci and the co-localization of H3K4me3 and H3K27ac at the *SAMMSON* TSS in Mel501. **g**, ChIP-qPCR of SOX10 in SK-MEL-28 at the indicated loci. The IgG antibody was used as a negative control. **h**, Gene expression levels in untreated (Mock) SK-MEL-28 cells or cells transfected with a control siRNA pool (siCtrl) or pools targeting MITF (siMITF) or SOX10 (siSOX10). Data are an average of three biological replicates ± standard error of the mean (s.e.m.). *r*, Pearson correlation coefficient.

data indicate that the lineage-specific expression profile of *SAMMSON* is a consequence of focal gene amplification and/or SOX10-mediated transcription.

Silencing *SAMMSON* using different locked nucleic acid (LNA)-modified antisense oligonucleotides (GapmeRs), which trigger RNase-H-mediated degradation of the target, greatly reduced the clonogenicity of all *SAMMSON*-expressing melanoma cultures independently of their *BRAF*, *NRAS* or *TP53* status (Fig. 2a, b and Extended Data Fig. 3a). The growth of 'invasive' melanoma cells (that is, MM165), and thereby intrinsically resistant to mitogen-activated protein kinase (MAPK) inhibitors[3] was also inhibited upon *SAMMSON* knockdown. Chronic treatment with inhibitors of BRAF(V600E) is invariably associated with the development of drug resistance[6]. Importantly, cells that acquired resistance to a BRAF(V600E) inhibitor (that is, SK-MEL-28-R) remained sensitive to *SAMMSON* targeting.

Compared with GapmeR11, GapmeR3 was more efficient in knocking down *SAMMSON* expression and had a greater growth inhibitory effect (Fig. 2a, b and Extended Data Fig. 3a, b). Although less efficient than GapmeR-mediated silencing, RNA interference (RNAi)-dependent knockdown of *SAMMSON* also inhibited melanoma growth (Extended Data Fig. 3c, d). In contrast, transfection of *SAMMSON*-targeting GapmeRs or short interfering RNA (siRNA) in MM001, NHME or non-melanoma cancer cell lines (that is, HCT116), all of which lack *SAMMSON*, did not cause growth inhibition (Fig. 2a, Extended Data Fig. 3b and data not shown). Moreover, forced expression of *SAMMSON* and *SAMMSON* mutants, carrying mismatches into the GapmeR target sequences, efficiently rescued growth inhibition induced by GapmeR3/11 (Extended Data Fig. 3e). Importantly, the latter experiment also indicated that *SAMMSON* exerts its pro-survival function in *trans*.

Flow cytometry and Caspase-Glo 3/7 assays indicated that this growth inhibition is, at least partly, due to a significant induction of the intrinsic mitochondrial apoptotic pathway (Fig. 2c and Extended Data Fig. 3b). Exposure to an inhibitor of cytochrome-*c*-induced pro-caspase-9 activation rescued *SAMMSON* knockdown-mediated growth inhibition (Extended Data Fig. 3f).

Notably, ectopic expression of *SAMMSON* in MM001 cells conferred them with a growth advantage (Fig. 2d, e and Extended Data Fig. 3g, h), confirming that *SAMMSON* exerts its pro-oncogenic function in *trans*.

Overcoming intrinsic and acquired resistance to MAPK inhibitors requires the simultaneous targeting of multiple pathways. Interestingly, GapmeR3 enhanced the cytotoxic effects of the BRAF(V600E) inhibitor vemurafenib and MEK inhibitor pimasertib (Fig. 2f and Extended Data Fig. 3i).

Knockdown of *SAMMSON* did not decrease MITF levels (Extended Data Fig. 4), ruling out the possibility that *SAMMSON* promotes melanoma survival by enhancing *MITF* transcription in *cis*. Instead, consistent with a function in *trans*, *SAMMSON* primarily localizes to the cytoplasm, with a fraction co-localizing with mitochondria (Extended Data Figs 5 and 6).

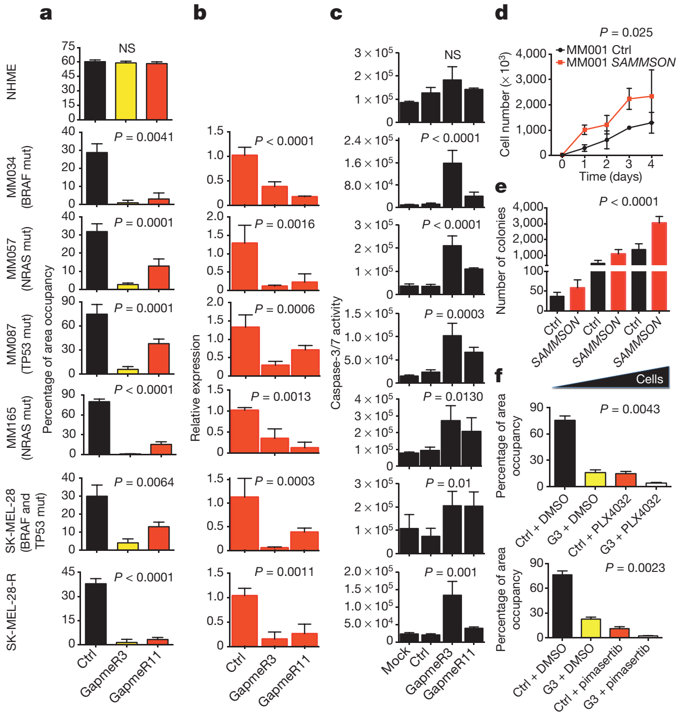


**Figure 2 | *SAMMSON* is required for melanoma growth and survival.** **a**, Quantification of colony formation assays 7 days after seeding of melanoma cells presented as the mean density (percentage of area occupancy) of three different biological replicates ± s.e.m. *P* values were calculated by ANOVA. mut, mutation **b**, Relative *SAMMSON* expression in cells transfected with a GapmeR control (Ctrl), GapmeR3 and GapmeR11. The results are the average of three different biological replicates ± s.e.m. *P* values were calculated by ANOVA. **c**, Caspase-Glo 3/7 assays in untreated cells (Mock) and cells transfected with GapmeR control (Ctrl), GapmeR3 and GapmeR11 (48 h post-transfection). The data are presented as the means of three different biological replicates ± s.e.m. *P* values were calculated by ANOVA. NS, not significant. **d**, Cell proliferation assays in MM001 upon ectopic *SAMMSON* expression. **e**, Quantification of colony formation assays 7 days after seeding 1,000, 5,000 or 10,000 cells described in **d**. The data are presented as the mean number of colonies of three different biological replicates ± s.e.m. at each density. *P* values were calculated by ANOVA. **f**, Quantification of colony formation assays 7 days after seeding of SK-MEL-28 cells transfected with a control GapmeR (Ctrl) or GapmeR3 and exposed to either vehicle or a half-maximum effective concentration ($EC_{50}$) dose of vemurafenib (PLX4032) or pimasertib. The data are presented as the mean density of three different biological replicates ± s.e.m. *P* values were calculated by ANOVA.

We next purified endogenous *SAMMSON* RNA complexes by adapting the RNA antisense purification–mass spectrometry (RAP–MS) and chromatin isolation by RNA purification (ChIRP)-like-MS methodologies[7,8]. After confirming that *SAMMSON* RNA, but not the housekeeping *TBC* and *UBC* messenger RNAs, was selectively retrieved (Fig. 3a), the *SAMMSON*-associated proteins were identified by mass spectrometry. Eighteen proteins were identified in multiple *SAMMSON* purifications from three independent biological samples (Extended Data Fig. 7a and Supplementary Tables 1, 2). The majority of these (12/18) are RNA-binding proteins, including XRN2, a protein involved in several key aspects of RNA metabolism. XRN2 was previously retrieved in ChIRP-MS experiments performed with different lncRNAs[8], indicating that XRN2 may be a bona fide lncRNAs interactor.

Among the enriched proteins, p32 attracted our attention because of its established role in mitochondrial metabolism. In addition, p32 expression is elevated in cancers[9–11] and p32 knockdown decreases the growth of various cancer cell lines, including melanoma[10,12]. Consistently, p32 levels were elevated in the MM cultures compared with NHME cultures (Extended Data Fig. 7b, c). RAP followed by

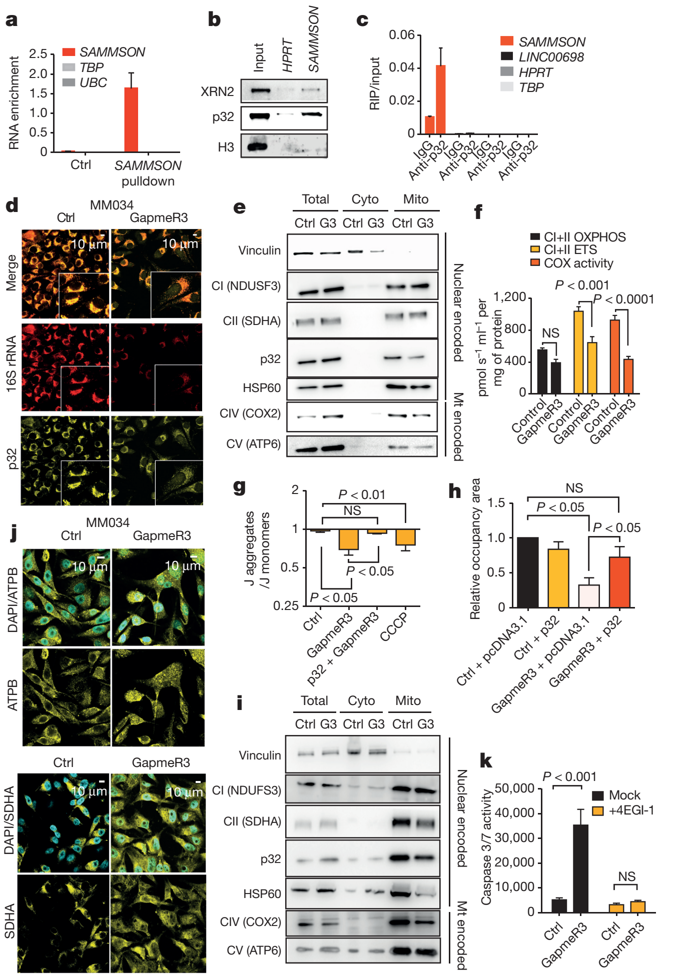


**Figure 3 | *SAMMSON* interacts with p32 to increase its mitochondrial localization and function.** **a**, *SAMMSON* (but not *TBP* or *UBC*) is specifically recovered by RAP. **b**, *SAMMSON* (and *HRPT*) pulldown after ultraviolet crosslinking and western blotting. **c**, *SAMMSON* is recovered by RIP using p32-specific antibodies. **d**, p32 immunofluorescence (yellow) and 16S rRNA fluorescence *in situ* hybridization (FISH; red) in MM034 cells transfected with a control GapmeR (Ctrl) or with GapmeR3. **e**, Western blotting 24 h after transfection of a control GapmeR (Ctrl) and GapmeR3 (G3). Cyto, cytosolic extracts; Mito, mitochondrial extracts; Mt encoded, mitochondria encoded. **f**, Measurement of the OXPHOS capacity (CI+II OXPHOS), electron transfer capacity of the respiratory chain (CI+II ETS) and COX activity using high-resolution respirometry of digitonin-permeabilized cells. Raw data were normalized to the total amount of proteins. The graph represents an average of four biological replicates ± s.e.m. *P* values were calculated by ANOVA. NS, not significant. **g**, Mitochondrial membrane potential (JC-1) in SK-MEL-28 cells transfected with a control GapmeR (Ctrl) or with GapmeR3, and either with an empty or a p32-expressing vector. The graph shows an average of four different biological replicates ± s.e.m. *P* values were calculated by ANOVA. **h**, Quantification of colony formation assays 5 days after seeding SK-MEL-28 cells transfected with a control GapmeR (Ctrl) or GapmeR3, and either with an empty or a p32-expressing vector. The data represent the occupancy area relative to Ctrl + pcDNA3.1. The data are an average of four biological replicates ± s.e.m. *P* values were calculated by ANOVA. **i**, Western blotting analysis 33 h after transfection of a control GapmeR (Ctrl) and GapmeR3 (G3). **j**, Immunofluorescence for SDHA and ATPB (and nuclear counterstaining using 4′,6-diamidino-2-phenylindole (DAPI)) in MM034 cells transfected with a control GapmeR (Ctrl) or with GapmeR3. **k**, Caspase-Glo 3/7 activity (72 h post-transfection) in SK-MEL-28 cells transfected with a control GapmeR (Ctrl) or GapmeR3, and incubated (8 h post-transfection) with a CAP-dependent translation inhibitor (4EGI-1). The graph shows an average of three biological replicates ± s.e.m. *P* value was calculated by ANOVA. **b**, **e**, **i**, For gel source data, see Supplementary Fig. 1.

western blotting analyses confirmed the association between p32 and *SAMMSON*. p32 was enriched in the *SAMMSON* pulldown relative to the samples bound to the negative control *HPRT* (Extended Data Fig. 7d). No enrichment was observed after purifications with either beads alone, in lysate pre-treated with RNase A, or using *SAMMSON*-null cells (Extended Data Fig. 7d and data not shown). To ensure that the interaction between *SAMMSON* and p32 is direct, we performed RAP followed by western blotting in cells in which covalent bonds between directly interacting RNA and proteins are created by ultraviolet crosslinking. p32 was still enriched in these conditions (Fig. 3b). We also performed RNA-immunoprecipitation (RIP) assays using p32-specific antibodies. Compared with the immunoglobulin G (IgG)-bound sample, all antibody-bound complexes showed a significant increase in the amount of *SAMMSON* RNA, but not of unrelated RNAs such as *HPRT*, *TBP* or *LINC00698* (Fig. 3c and Extended Data Fig. 7e).

Importantly, although no effect was seen on total p32 levels, a rapid decrease in its mitochondrial fraction, accompanied by an increase in nuclear targeting, was observed upon *SAMMSON* silencing (Fig. 3d, e and Extended Data Fig. 8a). p32 is required for the maturation of mitochondrial 16S rRNA, and thereby for the expression of mitochondrially encoded polypeptides, the maintenance of mitochondrial membrane potential and oxidative phosphorylation (OXPHOS)[10,13–16]. Consistently, the decrease in mitochondrial p32 was accompanied by reduced levels of 16S ribosomal RNA (Fig. 3d) and mitochondrial-DNA-encoded (COX2 and ATP6), but not nuclear-DNA-encoded (SDHA or NDUFS3), respiratory chain complex components (Fig. 3e). Moreover, the enzymatic activity of respiratory complexes I and IV, which contain proteins translated on mitochondrial ribosomes, was decreased upon *SAMMSON* silencing (Fig. 3f and Extended Data Fig. 8b). This decrease occurred before outer membrane permeabilization, as evidenced by the oxygraph profiles, and before induction of caspase-3/7 activity (data not shown).

The mitochondrial respiratory chain generates a proton gradient that establishes the mitochondrial membrane potential used to drive ATP production. A marked decrease in mitochondrial membrane potential was observed in *SAMMSON*-knockdown cells (Fig. 3g and Extended Data Fig. 8c), before the onset of apoptosis, accompanied by a slight, but significant, decrease in intracellular ATP levels (Extended Data Fig. 8d).

p32 controls mitochondrial homeostasis and integrity[15,17,18]. Accordingly, p32 silencing caused aberrant mitochondrial structures with fewer and fragmented cristae and reduced mitochondria matrix density. These effects were phenocopied upon *SAMMSON* knockdown (Extended Data Fig. 9).

Importantly, reintroduction of p32, but not of an N-terminally tagged version that can no longer be imported into the mitochondria, significantly rescued the *SAMMSON*-knockdown-dependent defect in mitochondrial membrane potential and growth inhibition (Fig. 3g, h and Extended Data Fig. 8e–g). *SAMMSON* silencing therefore inhibits melanoma survival at least partly by disrupting vital p32-mediated mitochondrial functions.

There is increasing evidence that mitochondrial translational deficiency induces a cytosolic stress response, which impairs cell growth[19]. This retrograde signalling is induced by the collapse of the mitochondrial membrane potential, which is critical for the import of nuclear-encoded proteins into the matrix[20]. This leads to the 'toxic' over-accumulation of mitochondrial precursors in the cytosol and activation of a stress response, which in yeast has been referred to as mitochondria precursor over-accumulation stress (mPOS)[21,22]. In turn, mPOS triggers cell cycle checkpoints and/or induces cell death depending on the cellular context. Interestingly, several hours after the collapse in the membrane potential, a decrease in the mitochondrial levels of several nuclear-encoded proteins (for example, SDHA, HSP60) was observed in *SAMMSON*-knockdown cells (Fig. 3i). Moreover, nuclear-encoded proteins such as SDHA and ATPB accumulated in the cytosol (and even in the nucleus for ATPB) of *SAMMSON*-knockdown cells (Fig. 3j and Extended Data Fig. 8h), indicating that *SAMMSON* silencing may lead to mitochondrial import defects, which, in turn, could activate mPOS.

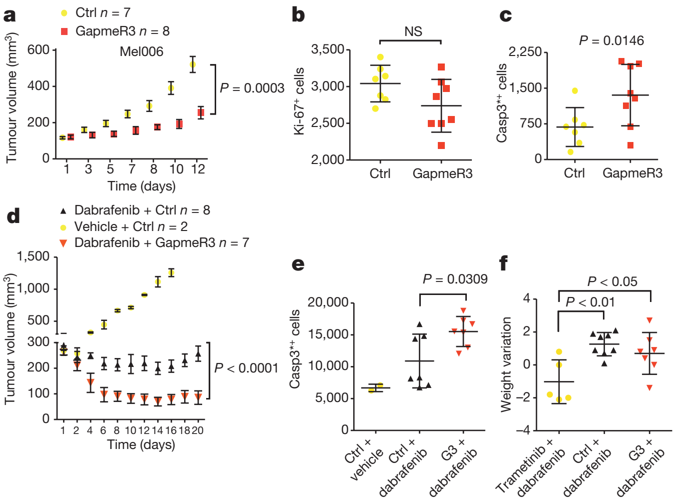


**Figure 4 | Therapeutic potential of *SAMMSON* targeting *in vivo*.** **a**, Tumour volume of cohorts of Mel006 PDX mice treated (intravenous injections) with a control GapmeR (Ctrl) or GapmeR3. Data are the means ± s.d. of different biological replicates (*P* value was calculated by two-ways ANOVA). **b**, **c**, Quantification of Ki-67-positive cells (**b**) and cleaved caspase 3-positive (Casp3*+; **c**) of melanoma lesions treated as described in **a**. *P* value was calculated by *t*-test. NS, not significant. **d**, Tumour volume of cohorts of PDX mice (Mel006) treated with combinations of control GapmeR (Ctrl) and GapmeR3 with either vehicle or dabrafenib by daily oral gavage (vehicle or dabrafenib) and intravenous injection of the GapmeRs every 2 days. Data are means ± s.d. of different biological replicates (*P* value was calculated by *t*-test). **e**, Quantification of cleaved caspase 3 of melanoma lesions treated as described in **d**. **f**, Weight variation of mice treated as described in **d** and mice exposed daily to dabrafenib and trametinib. *P* values were calculated by ANOVA.

In yeast, mPOS-induced cell death can be attenuated by reducing cytosolic protein translation, thus decreasing stress imposed by protein over-accumulation and aggregation[21]. Strikingly, exposure to a cap-dependent translation inhibitor significantly rescued *SAMMSON*-knockdown-induced apoptotic cell death (Fig. 3k). Notably, consistent with data indicating that mitochondrial translation stress may engage the p53 pathway[23], *SAMMSON* silencing triggered a significant p53 response in several melanoma cultures (data not shown). However, p53 knockdown did not rescue caspase-3/7 induction and growth inhibition, indicating that p53 is not strictly required for *SAMMSON*-knockdown-induced growth inhibition.

To test the therapeutic potential of *SAMMSON* targeting *in vivo*, we used two patient-derived xenograft (PDX) melanoma models (Mel006 and Mel010; Extended Data Fig. 10a). Intravenous treatment with the *SAMMSON*-targeting GapmeR3 significantly suppressed the growth of Mel006 tumours, decreased cell proliferation and increased apoptosis (Fig. 4a–c). Gene expression profiles of the GapmeR3-treated melanoma lesions showed enrichment for signatures associated with decreased cell proliferation, activation of p53 and decreased OXPHOS, mitochondrial ribosome biogenesis and respiratory chain complex activity (Extended Data Fig. 10b). Notably, this treatment did not cause any relevant adverse reaction or weight loss (data not shown). Tumour growth was also inhibited upon intra-tumour injections of GapmeR3 into Mel010 tumours (Extended Data Fig. 10c).

Importantly, whereas tumour growth was only inhibited after exposure to the BRAF(V600E) inhibitor dabrafenib alone, tumour regression and a significant increase in apoptosis were observed upon exposure to both dabrafenib and GapmeR3 (Fig. 4d, e and Extended Data Fig. 10d). Notably, these mice did not suffer from any relevant adverse events or weight loss, in contrast to mice treated with a combination of dabrafenib and MEK inhibitor (trametinib; Fig. 4f).

The observation that some cancer cells are dependent on OXPHOS for survival led to the development of agents that exploit bioenergetics and metabolic alterations in mitochondria[24]. The use of such agents is, however, complicated by their dangerous side effects. We show here that *SAMMSON* is a lineage-specific lincRNA that promotes melanoma survival through its ability to enhance the mitochondrial localization and function of p32, a protein required for the maintenance of OXPHOS[10]. Given that *SAMMSON* is expressed specifically in the vast majority (>90%) of melanomas, but not in normal adult tissues, these data identify *SAMMSON* as an attractive therapeutic target for the disruption of mitochondrial metabolism selectively in melanoma.

*MITF*, which is co-amplified with *SAMMSON* in a subset of melanomas, also promotes mitochondrial respiration by inducing PGC-1$\alpha$ (refs 25, 26). The *MITF-SAMMSON* amplicon therefore favours oxidative metabolism via two distinct mechanisms, the MITF–PGC-1$\alpha$ axis and the *SAMMSON*-p32 axis, making these cells particularly OXPHOS dependent.

Treatment of melanoma with BRAF inhibitors induces the MITF–PGC-1$\alpha$-dependent oxidative metabolic program and renders them highly dependent on OXPHOS[26]. This observation explains why co-targeting of *SAMMSON* and mutant BRAF promotes potent antitumour responses. *SAMMSON* targeting may therefore offer a novel therapeutic avenue to overcome the adaptive metabolic reprogramming that limits the efficacy of BRAF inhibitors. Moreover, as melanoma cells that acquire resistance to BRAF inhibitors remain addicted to *SAMMSON* expression, *SAMMSON* targeting might be a valid therapeutic approach to treat relapsed patients.

In addition, MITF-low invasive cells, although glycolytic, remain addicted to *SAMMSON* expression. These cells may still be dependent on mitochondria for functions other than ATP generation, such as fatty acid synthesis and glutaminolysis[27,28]. Interestingly, p32 has recently been shown to be required for MYC-induced addiction[11]. Through its ability to modulate mitochondrial protein synthesis, *SAMMSON* is also likely to affect processes outside the mitochondrion[19]. Consistently, we provide evidence that *SAMMSON* targeting decreases melanoma survival in an mPOS-dependent manner.

These findings warrant further investigation about the potential of *SAMMSON* as an informative biomarker of malignancy and as a highly selective and broad-spectrum anti-melanoma therapeutic target. Given the recent surge in optimism about antisense drugs, this therapeutic approach may be rapidly applicable to the clinic.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 21 December 2015; accepted 21 January 2016.**

1. Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* **436,** 117–122 (2005).
2. Gembarska, A. *et al.* MDM4 is a key therapeutic target in cutaneous melanoma. *Nature Med.* **18,** 1239–1247 (2012).
3. Verfaillie, A. *et al.* Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nature Commun.* **6,** 6683 (2015).
4. Harris, M. L., Baxter, L. L., Loftus, S. K. & Pavan, W. J. Sox proteins in melanocyte development and melanoma. *Pigment Cell Melanoma Res.* **23,** 496–513 (2010).
5. Laurette, P. *et al.* Transcription factor MITF and remodeller BRG1 define chromatin organisation at regulatory elements in melanoma cells. *eLife* **4,** 06857 (2015).
6. Flaherty, K. T. *et al.* Inhibition of mutated, activated BRAF in metastatic melanoma. *N. Engl. J. Med.* **363,** 809–819 (2010).
7. McHugh, C. A. *et al.* The *Xist* lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521,** 232–236 (2015).
8. Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* **161,** 404–416 (2015).
9. Amamoto, R. *et al.* Mitochondrial p32/C1QBP is highly expressed in prostate cancer and is associated with shorter prostate-specific antigen relapse time after radical prostatectomy. *Cancer Sci.* **102,** 639–647 (2011).
10. Fogal, V. *et al.* Mitochondrial p32 protein is a critical regulator of tumor metabolism via maintenance of oxidative phosphorylation. *Mol. Cell. Biol.* **30,** 1303–1318 (2010).
11. Fogal, V. *et al.* Mitochondrial p32 is upregulated in Myc expressing brain cancers and mediates glutamine addiction. *Oncotarget* **6,** 1157–1170 (2015).
12. Fogal, V., Zhang, L., Krajewski, S. & Ruoslahti, E. Mitochondrial/cell-surface protein p32/gC1qR as a molecular target in tumor cells and tumor stroma. *Cancer Res.* **68,** 7210–7218 (2008).
13. Muta, T., Kang, D., Kitajima, S., Fujiwara, T. & Hamasaki, N. p32 protein, a splicing factor 2-associated protein, is localized in mitochondrial matrix and is functionally important in maintaining oxidative phosphorylation. *J. Biol. Chem.* **272,** 24363–24370 (1997).
14. Yagi, M. *et al.* p32/gC1qR is indispensable for fetal development and mitochondrial translation: importance of its RNA-binding ability. *Nucleic Acids Res.* **40,** 9717–9737 (2012).
15. Hu, M. *et al.* p32 protein levels are integral to mitochondrial and endoplasmic reticulum morphology, cell metabolism and survival. *Biochem. J.* **453,** 381–391 (2013).
16. Matos, P. *et al.* A role for the mitochondrial-associated protein p32 in regulation of trophoblast proliferation. *Mol. Hum. Reprod.* **20,** 745–755 (2014).
17. Li, Y., Wan, O. W., Xie, W. & Chung, K. K. K. p32 regulates mitochondrial morphology and dynamics through parkin. *Neuroscience* **199,** 346–358 (2011).
18. Jiao, H. *et al.* Chaperone-like protein p32 regulates ULK1 stability and autophagy. *Cell Death Differ.* **22,** 1812–1823 (2015).
19. Richter-Dennerlein, R., Dennerlein, S. & Rehling, P. Integrating mitochondrial translation into the cellular context. *Nature Rev. Mol. Cell Biol.* **16,** 586–592 (2015).
20. Chacinska, A., Koehler, C. M., Milenkovic, D., Lithgow, T. & Pfanner, N. Importing mitochondrial proteins: machineries and mechanisms. *Cell* **138,** 628–644 (2009).
21. Wang, X. & Chen, X. J. A cytosolic network suppressing mitochondria-mediated proteostatic stress and cell death. *Nature* **524,** 481–484 (2015).
22. Wrobel, L. *et al.* Mistargeted mitochondrial proteins activate a proteostatic response in the cytosol. *Nature* **524,** 485–488 (2015).
23. Richter, U. *et al.* A mitochondrial ribosomal and RNA decay pathway blocks cell proliferation. *Curr. Biol.* **23,** 535–541 (2013).
24. Fantin, V. R. & Leder, P. Mitochondriotoxic compounds for cancer therapy. *Oncogene* **25,** 4787–4797 (2006).
25. Vazquez, F. *et al.* PGC1$\alpha$ expression defines a subset of human melanoma tumors with increased mitochondrial capacity and resistance to oxidative stress. *Cancer Cell* **23,** 287–301 (2013).
26. Haq, R. *et al.* Oncogenic BRAF regulates oxidative metabolism via PGC1$\alpha$ and MITF. *Cancer Cell* **23,** 302–315 (2013).
27. Wise, D. R. & Thompson, C. B. Glutamine addiction: a new therapeutic target in cancer. *Trends Biochem. Sci.* **35,** 427–433 (2010).
28. Dang, C. V. Links between metabolism and cancer. *Genes Dev.* **26,** 877–890 (2012).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** GapmeRs were designed by J. Lai. This work was supported by the Fonds Wetenschappelijk Onderzoek (FWO; #G.0646.14N and #3G056613), Foundation Against Cancer (STK#2014-126), UGent-IOF (F2013/IOF-Advanced/676) and STK grant F/2014/376. The PDX studies were funded by GOA/14/012 (KULeuven) and KPC_29_005 (Belgian Ministries of Health). The authors wish to thank N. Samyn for help with MS, M. Van Gele for providing NHME cultures and P. Wolter for scientific discussions and his central role in the establishment of the melanoma PDX platform. E.L. is a recipient of a postdoctoral fellowship from the Marie-Curie/VIB OMICS program. M.S. is the recipient of EMBO fellowship (ALTF 648-2013). F.A. is a senior researcher from the Research Fund Flanders (FWO). P.M. and R.V. are recipients of FWO postdoctoral and PhD fellowships, respectively. D.L. is supported by the Fonds National de la Recherche (FRS/FNRS). I.D. is supported by the Institut National du Cancer PAIR melanoma (MELA13-002), the "France Génomique" consortium (ANR10-INBS-09-08), and ANR-10-LABX-0030-INRT. The I.D. laboratory is an "équipe labellisée Ligue Nationale contre le Cancer".

**Author Contributions** E.L. and R.V. performed most experiments. M.S. performed the experiments described in Fig.3f. P.L., S.A. and I.D. provided ChIP-seq data. J.W. and J.v.d.O. provided melanoma clinical samples. J.v.d.O. and E.R. provided histopathology support. S.E. and K.G. performed the mass spectrometric measurement and analysis. C.L., K.V., S.V.A. and P.M. performed copy number variant analysis and profiling experiments and interpretation of the data. A.R., E.H. and F.A. provided support with the PDX models. P.B. provided technical support for the electron microscopy. P.M., M.F. and S.A. provided bioinformatics support. D.L.J.L., B.d.S., P.M. and J.V. helped with the interpretation of the data. J.-C.M. and E.L. designed most of the experiments and wrote the manuscript.

**Author Information** Mass spectrometry data have been deposited in the Proteomics Identifications Database under accession numbers PXD002565. Gene expression data have been deposited in the Gene Expression Omnibus under accession number GSE70180. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.-C.M. (JeanChristophe.Marine@cme.vib-kuleuven.be).

## METHODS

**Cell culture.** The melanoma cell lines Mel501, SK-MEL-28 (obtained from ATCC) were grown in 5% $CO_2$ at 37 °C in RPMI 1640-glutamax (Gibco, Invitrogen) or DMEM (for Mel501) supplemented with 10% FBS (Hyclone, Thermo Fisher Scientific). Sequences of the GapmeRs and siRNAs are indicated later.

Caspase 9 inhibitor was purchased from Merk Millipore (218761). 4EGI-1 was purchased from Selleckchem and used at a final concentration of 10 μM. The patient-derived low-passage MM cell lines (a gift from the Ghanem laboratory) were grown in F-10 (Gibco, Invitrogen), supplemented with 10% FBS (Hyclone, Thermo Fisher Scientific) and 12 mM glutamine.

NHMEs were grown in MGM-4 melanocyte growth medium (Lonza). Cells were transfected by Lipofectamine 2000 according to manufacturer instructions with 25 nM of GapmeR (Exiqon). For siRNA experiments the cells were transfected with 25–50 nM siRNAs or Dharmacon Pools and harvested 24 and/or 48 h after transfection. All cell lines used in this study were mycoplasma negative.

**Human melanoma samples.** Early stages of melanoma from the UZ Leuven archive were isolated by laser capture microdissection and RNA extracted with Arcturus PicoPure Frozen RNA Isolation Kit (Life Technologies).

**RAP–MS.** For affinity purification of *SAMMSON* protein targets, 100 μg of Streptavidin Sepharose High Performance (GE Healthcare) was coupled to 400 pmol of biotinylated probes against *SAMMSON* (Biosearch Technologies) overnight at 4 °C. Cells ($60 \times 10^6$ cells per sample) were lysed in 2 ml of pull-out buffer (20 mM Tris-HCl, pH 7.5, 200 mM NaCl, 2.5 mM $MgCl_2$, 0.05% Igepal, 60 U Superase-In per ml (Ambion), 1 mM dithiothreitol (DTT) and a cocktail of protease inhibitors) and incubated for 3 h with the beads at 4 °C. As a negative control, an additional sample was digested with 10 μg ml$^{-1}$ RNase A digestion for 10 min at room temperature before incubation with *SAMMSON* probes.

For the crosslinking experiments, cells were washed once in PBS, crosslinked dry at 400 mj cm$^{-2}$ and lysed. During the washes the amount of Triton X-100 was doubled and SDS was added to a final concentration of 0.02%.

For mass spectrometry, samples were processed by a short separation on SDS–PAGE gels (Biorad) to remove contaminants possibly interfering with downstream analysis. After excision, washing and drying of the gel band, digestion buffer (50 mM $NH_4HCO_3$, 5 μg ml$^{-1}$ trypsin) was added to fully submerge the dried gel band, and the sample was digested for 16 h at 37 °C. The generated peptide mixtures were acidified, dried and re-dissolved in a 2% $CH_3CN$ (acetonitrile), 0,1% formic acid solution.

The obtained peptide mixtures were introduced into a liquid chromatograpy-tandem mass spectrometry (LC—MS/MS) system through an ultimate 3000 RSLC nano LC (Thermo Scientific) inline connected to a Q Exactive mass spectrometer (Thermo Fisher Scientific). The sample mixture was first loaded on a trapping column (made in house, 100 μm internal diameter (i.d.) × 20 mm, 5 μm beads C18 Reprosil-HD, Dr. Maisch, Ammerbuch-Entringen). After flushing from the trapping column, the sample was loaded on an analytical column (made in house, 75 μm i.d. × 150 mm, 5 μm beads C18 Reprosil-HD, Dr. Maisch) packed in the nanospray needle (PicoFrit SELF/P PicoTip emitter, PF360-75-15-N-5, NewObjective). Peptides were loaded with loading solvent (0.1% TFA in water) and separated with a linear gradient from 98% solvent A′ (0.1% formic acid in water) to 40% solvent B′ (0.08% formic acid in water/acetonitrile, 20/80 (v/v)) in 30 min at a flow rate of 300 nl min$^{-1}$. This was followed by a 15 min wash reaching 99% solvent B′. The mass spectrometer was operated in data-dependent, positive ionization mode, automatically switching between MS and MS/MS acquisition for the 10 most abundant peaks in a given MS spectrum.

The source voltage was 3.4 kV, and the capillary temperature was 275 °C. One MS1 scan ($m/z$ 400—2,000, AGC target $3 \times 10^6$ ions, maximum ion injection time 80 ms) acquired at a resolution of 70,000 (at 200 $m/z$) was followed by up to 10 tandem MS scans (resolution 17,500 at 200 $m/z$) of the most intense ions fulfilling the defined selection criteria (AGC target $5 \times 10^4$ ions, maximum ion injection time 60 ms, isolation window 2 Da, fixed first mass 140 $m/z$, spectrum data type: centroid, underfill ratio 2%, intensity threshold $1.7 \times 10^4$, exclusion of unassigned, 1, 5–8, and >8 charged precursors, peptide match preferred, exclude isotopes on, dynamic exclusion time 20 s). The HCD collision energy was set to 25% normalized collision energy and the polydimethylcyclosiloxane background ions at 445.120025 Da were used for internal calibration (lock mass).

From the MS/MS data in each LC run, Mascot Generic Files were created using Distiller software (version 2.4.3.3, Matrix Science, http://www.matrixscience.com/distiller.html). While generating these peak lists, grouping of spectra was allowed in Distiller with a maximal intermediate retention time of 30 s and a maximal intermediate scan count of 5 was used where possible. Grouping was done with 0.005 Da precursor ion tolerance. A peak list was only generated when the MS/MS spectrum contains more than 10 peaks. There was no de-isotoping and the relative signal to noise limit was set at 2. These peak lists were then searched using the Mascot search engine (MatrixScience)[29] with the Mascot Daemon interface (version 2.4.1, Matrix Science). Spectra were searched against the human protein entries in the Swiss-Prot database (SP2014_07; 20284 sequence entries). Variable modifications were set as methionine oxidation, pyro-glutamate formation of N-terminal glutamine, propionamide formation on cysteine and acetylation of the protein N terminus. The mass tolerance on precursor ions was set to 10 ppm (with Mascot's C13 option set to 1) and on fragment ions to 20 mmu. The instrument setting was put on ESI-QUAD. Enzyme was set to trypsin, allowing for one missed cleavage. Only peptides that were ranked first and scored above the threshold score, set at 99% confidence, were withheld. The protein candidates that were pursued in this work were consistently identified with at least two different peptides in the relevant conditions (no peptides detected in the bead controls).

**RIP.** RIP was performed as previously described[7]. p32 and XRN2 were immunoprecipitated using 4 μg of specific antibody (Bethyl laboratories) coupled to 50 μl of protein G Dynabeads (Invitrogen) for 3 h.

**Cellular fractionation.** Briefly, total nuclear and cytoplasmic extracts were isolated from 20-cm ø dishes using Nuclei EZ prep (Sigma-Aldrich) according to the manufacturer's instructions. RT–qPCR for *MALAT1* and *TBP* were used to assess the purity of the fractions.

Mitochondria were purified from 20-cm ø dishes using mitochondria isolation kit for cultured cells (Thermo Fisher Scientific). Mitochondria enrichment was validated by western blot using antibodies directed against mitochondrial proteins (that is, ATPB). Mitoplasts were obtained by incubating mitochondria in hypotonic buffer (HEPES pH 7.2) for 20 min on ice.

**Antibodies.** Western blotting experiments were performed using the following primary antibodies: vinculin (V9131, Sigma-Aldrich, 1:10,000), histone 3 (ab1791, Abcam, 1:1,000), GAPDH (ab9485, Abcam, 1:1,000), p32 (A302-863A, Bethyl Laboratories, 1:5,000), XRN2 (A301-103A, Bethyl Laboratories, 1:2,000), SOX10 (N-20, Santa Cruz, 1:500), MITF (ab12039, Abcam, 1:1,000) and SDHA (Abcam AB14715, 1:1,000), COX2 (molecular probes A6404, 1:10,000), HSP60 (BD 611562, 1:5,000), NDUFS3 (Abcam 14711, 1:1,000), ATP6 (Abcam 192423, 1:1,000).

**Cell growth and cell death assays.** To detect cell death, cells were stained for 15 min with annexin V and PI using the FITC Annexin V Apoptosis Detection Kit II (BD Biosciences) according to the manufacturer's instructions. Cell death was detected on a MACSQuant VYB (Miltenyi Biotech BV) and data were analysed with FlowJo software (Tree Star).

Caspase 3 and 7 activity was measured using Caspase-Glo 3/7 luciferase assay (Promega) and VICTOR X4 Reader (PerkinElmer) 48 and 72 h after transfection of GapmeRs.

**Colony assay and cell count.** Cells were plated in six-well plates at the appropriate density and cultured for 1 week. The cells were washed with PBS, fixed and stained for 15 min with a 1% crystal violet in 35% methanol solution.

For vital counts, cells were stained with Trypan Blue (Sigma-Aldrich) and counted with TC20 automated cell counter (Biorad).

BRAF and MEK inhibitors (vemurafenib and pisertimab) were used at a concentration of 5 and 1 μM, respectively.

***SAMMSON* cloning, mutagenesis and lentiviral transduction.** *SAMMSON* was synthesized by GenScript and cloned into pPGK (Addgene) lentiviral vector. Lentivirus produced in HEK293 cells were used to infect the MM001 cells. Successfully infected cells were selected in puromycin (0.5 μg ml$^{-1}$)-containing medium for 1 week.

**Pharmacological treatment of mice.** The Mel0010 and Mel006 PDX models derived from two different metastatic melanoma lesions, both carrying the BRAF(V600E) mutation. Written informed consent was obtained from both patients and all procedures involving human samples were approved by the UZ Leuven/KU Leuven Medical Ethical Committee (# ML8713/S54185). All procedures involving animals (NMRI nude, 4-week-old females) were performed in accordance with the guidelines of the Catholic University of Leuven (KU Leuven) Animal Care and Use Ethical Committee (P147/2012).

Once tumours reached 150 mm$^3$, 10 mg kg$^{-1}$ of GapmeR3 or control GapmeR were injected i.v. or directly into the tumours every 2 days for up to 20 days. For combination therapy with BRAF inhibitor, cohorts of Mel006 were enrolled into the experiment once tumours reached 250 mm$^3$ in volume. Dabrafenib or vehicles were administrated daily by oral gavage. The GapmeRs, at a concentration of 10 mg kg$^{-1}$, were administrated i.v. every 2 days.

No specific randomization method was used. According to animal welfare guidelines, mice have to be killed when tumours reach a volume of 2,000 mm$^3$ or when their body weight decreases more than 20% from the initial weight. Mice used in this paper never reached or overcame these limits. The investigators were blinded for the evaluation of the results.

**FISH for *SAMMSON*.** For detection of *SAMMSON* at a single-cell level, a pool of 48 FISH probes was designed using the Stellaris probe designer software (Biosearch

Technologies). Cells were grown on slides and fixed in 3.7% formahldeyde and permeabilized in ethanol 70%. Hybridization was carried out overnight at 37 °C in 2× SSC, 10% formamide and 10% dextran. Cells were counterstained with DAPI and visualized using an Olimpus Fluoview FV1200 using a LD635 laser for Cy5, LD559 HeNe for Cy3.5 and LD405 for DAPI.

**Electron microscopy.** Cells were fixed with 2.0% paraformaldehyde/2.5% EM-grade glutaraldehyde in 0.1 M sodium cacodylate buffer (pH 7.4) at 37 °C overnight and collected. After fixation, samples were placed in 1% osmium tetroxide for 2 h and dehydrated in a graded series of ethyl alcohol. The agar-embedded samples were cut with a Leica UCT ultramicrotome in 50–70 nm sections and imaged in a JEOL JEM1400 transmission electron microscope at 80 kV.

**Immunohistochemistry and immunofluorescence.** Tumour biopsies formalin-fixed and paraffin embedded from the UZ Leuven archives, were cut in sections of about 4 μm. Samples were deparaffinized and dehydrated with xylene and graded alcohols, and subsequently rehydrated with demineralized water. Specimens were stained with haematoxylin and eosin and immunohistochemistry was performed using microwave pre-treatment of slides for antigen retrieval. Antibodies against Ki-67 (SP6, Thermo Fisher Scientific #RM-9106-S, clone SP6, 1:200) and cleaved caspase 3 (Asp175, Cell Signaling Technology, 1:300) were applied, in conjunction with goat anti-rabbit horseradish peroxidase (HRP)-conjugated antibodies (DAKO) and visualized by DAB reaction. To evaluate the stainings, positive cells in the blue channel were counted in four different fields using ImageJ. Statistical significance was calculated by two-way ANOVA.

For immunofluorescence, cells were grown on slides and permeabilized in Triton X-100-containing buffer. Blocking was performed in 5% BSA and 10% goat serum. To detect p32 a rabbit antibody from Bethyl Laboratories was used 1:1,000 followed by staining with an anti-rabbit AlexaFluor-488 (Life Technologies, 1:500). Images where analysed on an Olympus fluoview FW1200 using a LD635 laser for Cy5, LD559 HeNe for Cy3.5 and LD405 for DAPI. To detect ATPB (AB14730) and SDHA (ab66484) antibodies from Abcam were used according to the manufacturer's instructions.

**RT–qPCR lncRNA expression profiling.** Expression of 1,718 lncRNAs was measured on the NCI60 cell line panel using the SmartChip Human lncRNA1 panel (Wafergen Biosystems). NCI60 cell lines were obtained through the Developmental Therapeutics Program (National Cancer Institute (NCI)). Previously misclassified cell lines NCI/ADR-RES and MDA-MB-435 were correctly annotated in the analysis. RNA was isolated using the miRNeasy mini kit (Qiagen) and reverse transcribed (2 μg) using the iScript Advanced RT kit (Bio-Rad) according to the manufacturer's instructions. A qPCR reaction mixture containing 2 μg of cDNA and SsoAdvanced Universal SYBR Green mastermix was dispensed using the MultiSample NanoDispenser across the 5,184-nanowell SmartChip and analysed using the SmartChip Cycler (Wafergen Biosystems). lncRNA expression was measured in triplicate and median Cq values were normalized using the global mean normalization strategy[30]. RNA from normal adult tissues (Ambion and Biochain) and primary melanocytes was reverse transcribed using the iScript RT kit (Bio-Rad). *SAMMSON* expression was measured by qPCR on a LightCycler 480 (Roche) and normalized in qbase+ (Biogazelle) using *HPRT1*, *TBP* and *SDHA* as reference genes.

**Microarray gene expression profiling.** Protein-coding gene expression in xenografts was measured using a custom gene expression microarray (SurePrint 8 × 60k, Agilent) or commercial microarray (SurePrint G3 Human Gene Expression v.2, Agilent) respectively. RNA (100 ng) was labelled using the Quick Amp polyA labelling kit and hybridized according to the manufacturer's instructions. Slides were scanned using a high-resolution microarray scanner (Agilent) and probe intensities were extracted using Feature Extraction software (Agilent). Signals were background corrected and Quantile normalized using the limma package in R. Only those probes expressed twofold above the mean signal of a negative control probe were retained for further analysis.

**Differential gene expression and pathway analysis.** Probes were collapsed to gene level by retaining the probe with the highest average signal across all samples. Probes expressed in less than half of the samples were discarded for differential gene expression analysis. Genes differentially expressed across four cell lines (SK-MEL-28, MM034, MM057, MM087) upon knockdown of *SAMMSON* were identified using the limma package with cell line as a blocking variable. For differential expression analysis in xenograft tumours, no blocking variable was applied. Genes with an adjusted $P$ value $< 0.05$ and twofold expression change were selected as differentially expressed. Gene set enrichment analysis was performed on mRNA lists, pre-ranked according to the limma $t$-statistic using all curated gene sets from the Molecular Signatures Database.

**Generalized additive models.** Protein-coding gene expression data for the NCI60 cohort were obtained from the Developmental Therapeutics Program (NCI) data portal. Expression of each of the 1,718 lncRNAs (response variable) was analysed in relation to cancer type and transcription factor expression (predictor variables) using generalized additive models (GAMs). A total of 1,270 transcription factors were analysed. GAMs were constructed using the mgcv package in R using the following parameters: family = "Gaussian", link = "identity", method = "GCV.Cp". The GAM value for each model represents the percentage of lncRNA expression variance explained by the predictor variable.

**Quantification of RNA-FISH data.** Tiff Z-stacks were imported using the StarSearch software package for automated spot counting of defined regions. Nuclear and total cell spots were counted for MM057, MM087 and SK-MEL-28 cells. Nuclear lncRNA copy number was defined as the average number of nuclear spots while cytoplasmic lncRNA copy number was defined as the difference between the number of total cell spots and the number of nuclear spots.

**Analysis of expression and copy number data from the TCGA cohort.** Level 3 segmented DNA SNP array data for 386 melanoma tumours was obtained through the TCGA data portal. *SAMMSON* copy number was defined as the mean log ratio of the overlapping segment(s). Expression data for *SAMMSON* (uc003dog) and *SOX10* (uc003aun) in melanoma and 23 additional cancer types were extracted from level 3 TCGA RNA-seq data, totalling 8,085 primary tumour samples.

***SAMMSON* transcript boundaries.** The *SAMMSON* TSS site was identified by means of CAGE-seq tags in two melanoma cell lines obtained from the FANTOM5 study (http://fantom.gsc.riken.jp/5/). The *SAMMSON* 3′ end was identified by means of 3′ RACE. Briefly, an anchored oligo-dT primer (AATACGACTCACTATAGGCGCTTTTTTTTTTTTTTTTTTTTVN) was used for cDNA synthesis of SK-MEL-28 RNA using the iScript select reverse transcription kit (BioRad). The 3′ end was amplified (forward: AGCCAAATTTCAATGA GCCCCT; reverse: AATACGACTCACTATAGGC GC) and the resulting PCR amplicon was sized using the labchip (Calliper). The 3′ end was identified based on the amplicon length and location of the forward primer.

**High-resolution respirometry.** SK-MEL-28 cells were resuspended in 20 mM HEPES, 110 mM sucrose, 10 mM $KH_2PO_4$, 20 mM taurine, 60 mM lactobionic acid, 3 mM $MgCl_2$, 0.5 EGTA, pH 7.1, 1 mg ml$^{-1}$ fatty-acid-free BSA and catalase 280 U ml$^{-1}$ . Mitochondrial oxygen consumption was measured after plasma membrane permeabilization with digitonin 10 μg ml$^{-1}$ in an Oroboros 2k apparatus at 37 °C. The oxygen consumption rates, expressed as pmol $O_2$ s$^{-1}$ mg protein$^{-1}$, were measured after addition of the following substrates and specific inhibitors. (1) 2.5 mM pyruvate, 1 mM malate, 10 mM glutamate in the absence of ADP to determine complex-I-driven non phosphorylating respiration (CI leak). (2) 2.5 mM ADP to determine complex-I-driven phosphorylating respiration (CI OXPHOS). The coupling efficiency between oxygen consumption and phosphorylation was estimated as the ratio between CI OXPHOS and CI leak. (3) 10 mM succinate to determine the phosphorylating oxygen consumption driven by simultaneous activation of complex I and II (CI+II OXPHOS). (4) Titrating concentrations of the mitochondrial uncoupler CCCP to reach the maximal, uncoupled respiration (CI+II electron transfer system). (5) 0.5 μg ml$^{-1}$ antimycin A to block mitochondrial respiration at the level of complex III, and estimate residual non-mitochondrial oxygen consumption. (6) 2 mM ascorbate, 0.5 mM TMPD to measure cytochrome *c* oxidase (CIV)-driven respiration. (7) 10 μg ml$^{-1}$ cytochrome *c* to evaluate mitochondrial outer membrane damage. (8) 250 μM potassium cyanide to measure residual chemical background. CIV-driven respiration was measured as the cyanide sensitive oxygen consumption.

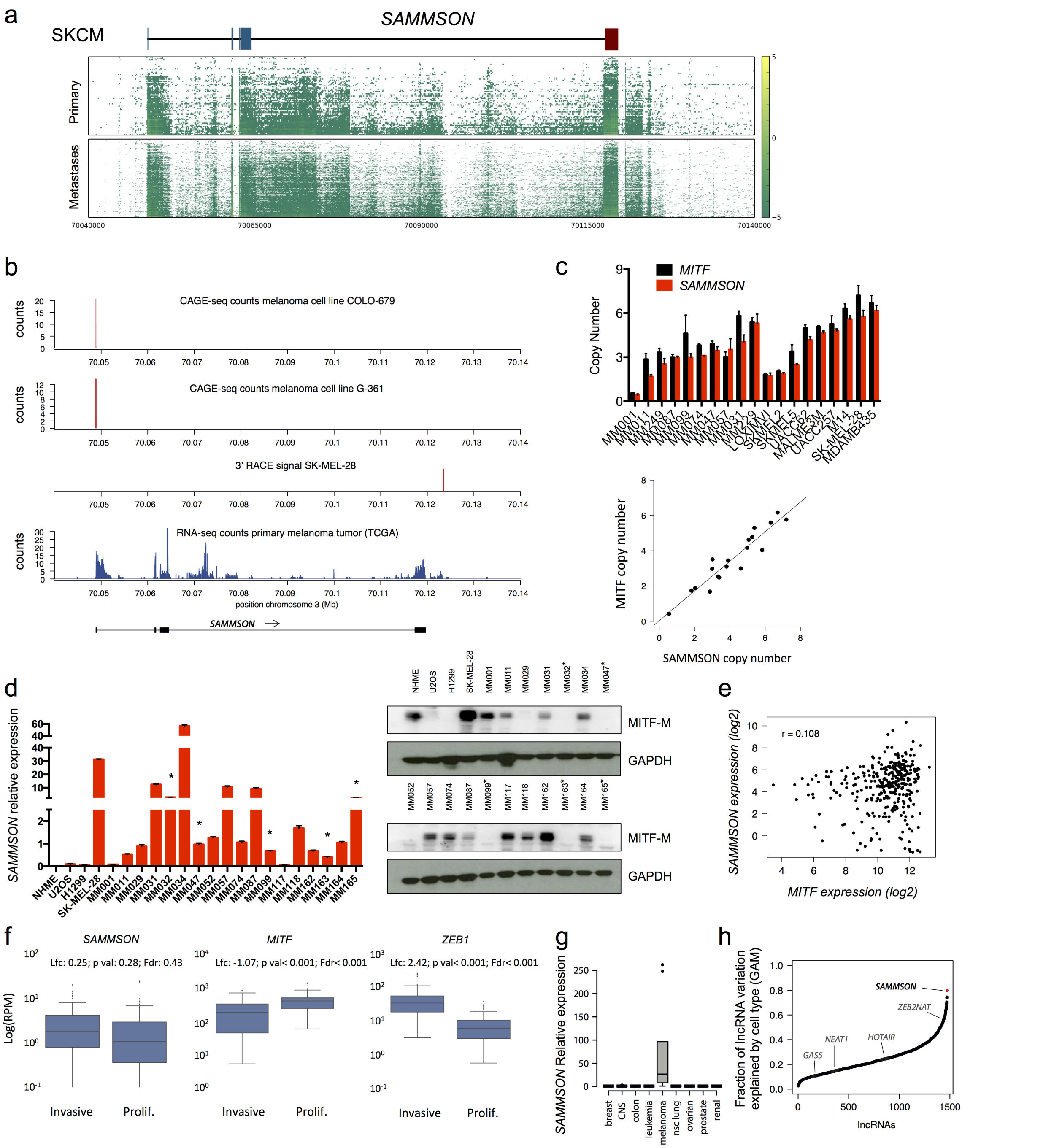
**Measurement of ATP production, mitochondrial membrane potential and ROS production.** Luciferase-based measurement of ATP production was obtained using Molecular Probes kit according to the manufacturer's instructions.

Mitochondrial membrane potential was measured by flow cytometry on a MACSQuant VYB (Miltenyi Biotech BV) using JC-1 dye (Molecular Probes). Data were analysed with FlowJo software (Tree Star) and the extent of depolarization was defined by the ratio between J-aggregates and J-monomers.

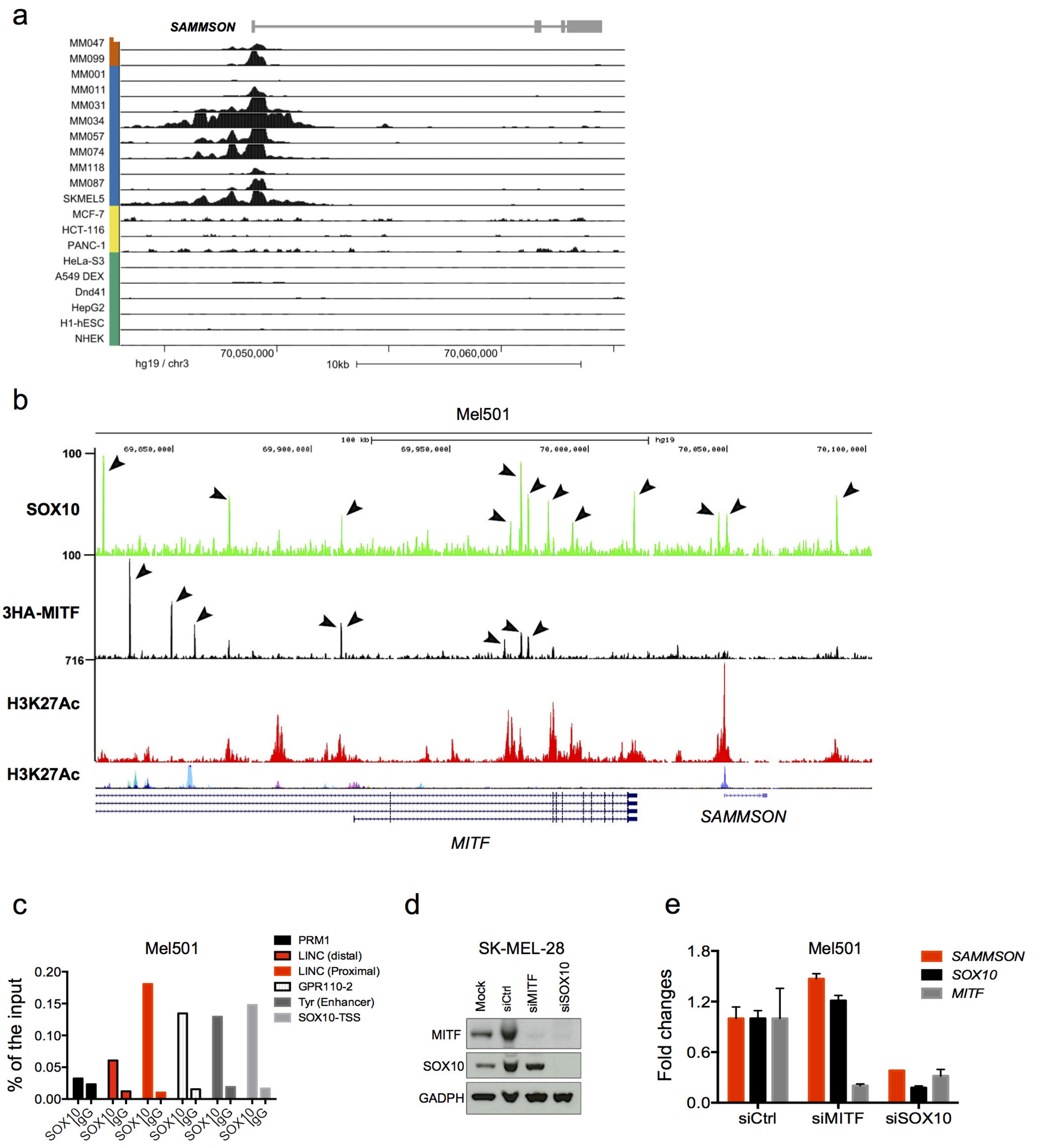
**Primers and siRNAs used.** *HPRT* forward, AGCCAGACTTTGTTGGA TTTG; reverse, TTTACTGGCGATGTCAATAAG; *TBP* forward, CGGC TGTTTAACTTCGCTTC; reverse, CACACGCCAAGAAACAGTGA; *UBC* forward, ATTTGGGTCGCGGTTCTTG; reverse, TGCCTTGACATT CTCGATGGT; *MITF-M* forward, CATTGTTATGCTGGAAATGCTAGAA; reverse, GGCTTGCTGTATGTGGTACTTGG; *SOX10* forward, TACCCGC ACCTGCACAAC; reverse, TTCAGCAGCCTCCAGAGC; *SOX9* forward, GCAAGCTCTGGAGACTTCTG; reverse, GTACTTGTAATCCGGGTGGTC; *p32* forward, ACACCGACGGAGACAAAG; reverse, GGGATGCTGTTG TTAATGTTG; *MALAT1* forward, GGATTCCAGGAAGGAGCGAG; reverse, ATTGCCGACCTCACGGATTT; *SAMMSON* forward, TTCCTCAACTATGCAACTCAA; reverse, TAGACTACGGGCTCATGACTT; *SAMMSON* forward #2, CCTCTAGATGTGTAAGGGTAGT; reverse #2, TTGAGTTGCATAGTTGAGGAA; *GPR110* forward, CAGTATTG

TGGCGGAAAAGC; reverse, CATCTTGCATGGCCCCA; *TYR* forward, AGCAGGCTCAGTCGATACAG; reverse, CACTGGGAATGAAGGGCAAG; *SAMMSON* GapmeR3, GTGTGAACTTGGCT; GapmeR11, TTTGAGAGTTGGAGGA; non-targeting GapmeR; TCATACTATATGACAG; sip32.1 sense, GGTTGAAGAACAGGAGCCT; antisense, AGGCTCCTGTTCTTCAACC; sip32.2 sense, TCACGGTCACTTTCAACAT; antisense, ATGTTGAAAGTGACCGTGA; *SAMMSON* siRNA sense, GUCGCUAGACAUUUGAGGA[dA][dA]; siRNA antisense UCCUCAAAUGUCUAGCGAC[dA] [dA]; SOX10 and MITF knockdown, Dharmacon Smartpool.

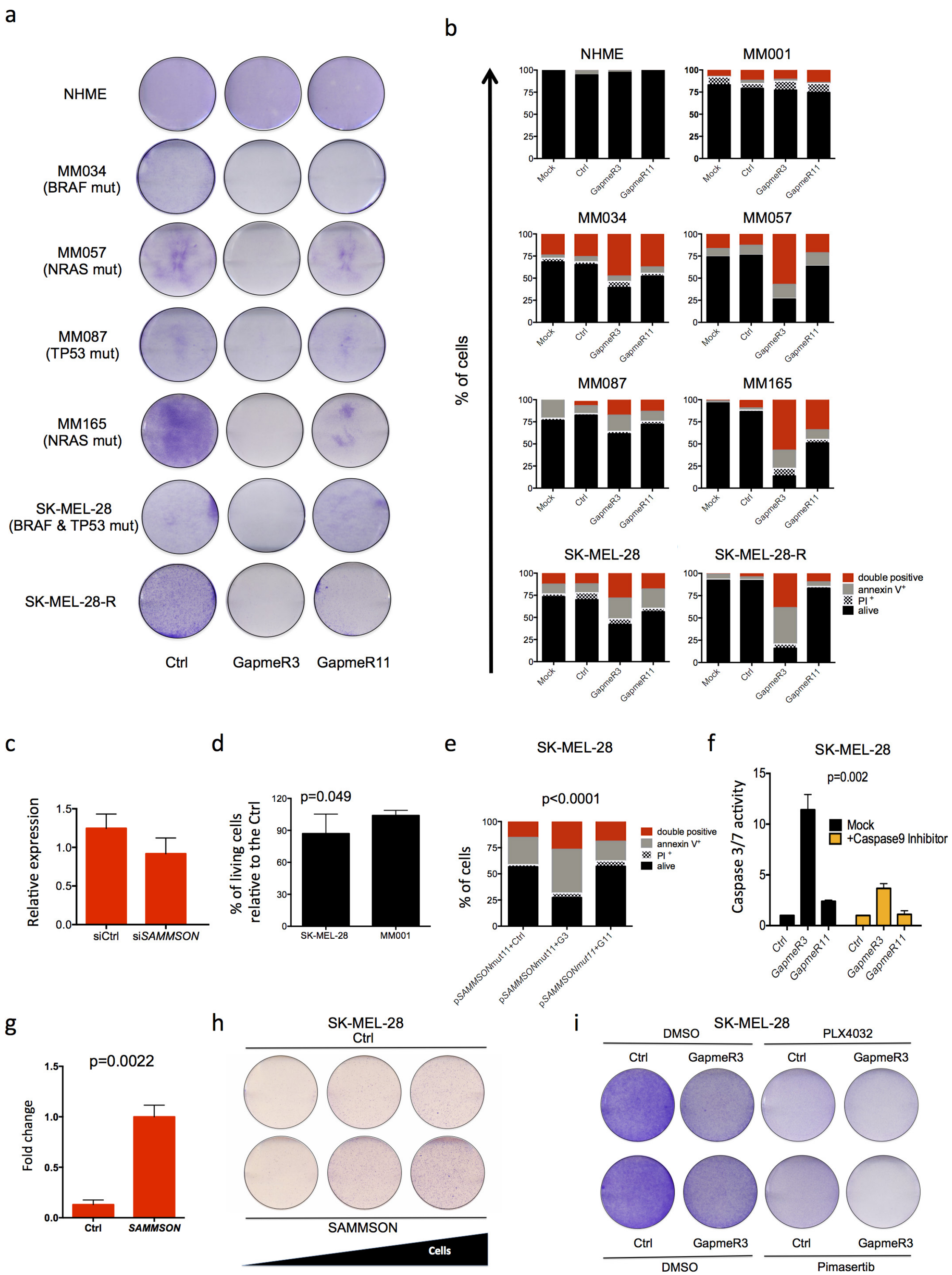
29. Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20,** 3551–3567 (1999).
30. Mestdagh, P. *et al.* A novel and universal method for microRNA RT–qPCR data normalization. *Genome Biol.* **10,** R64 (2009).



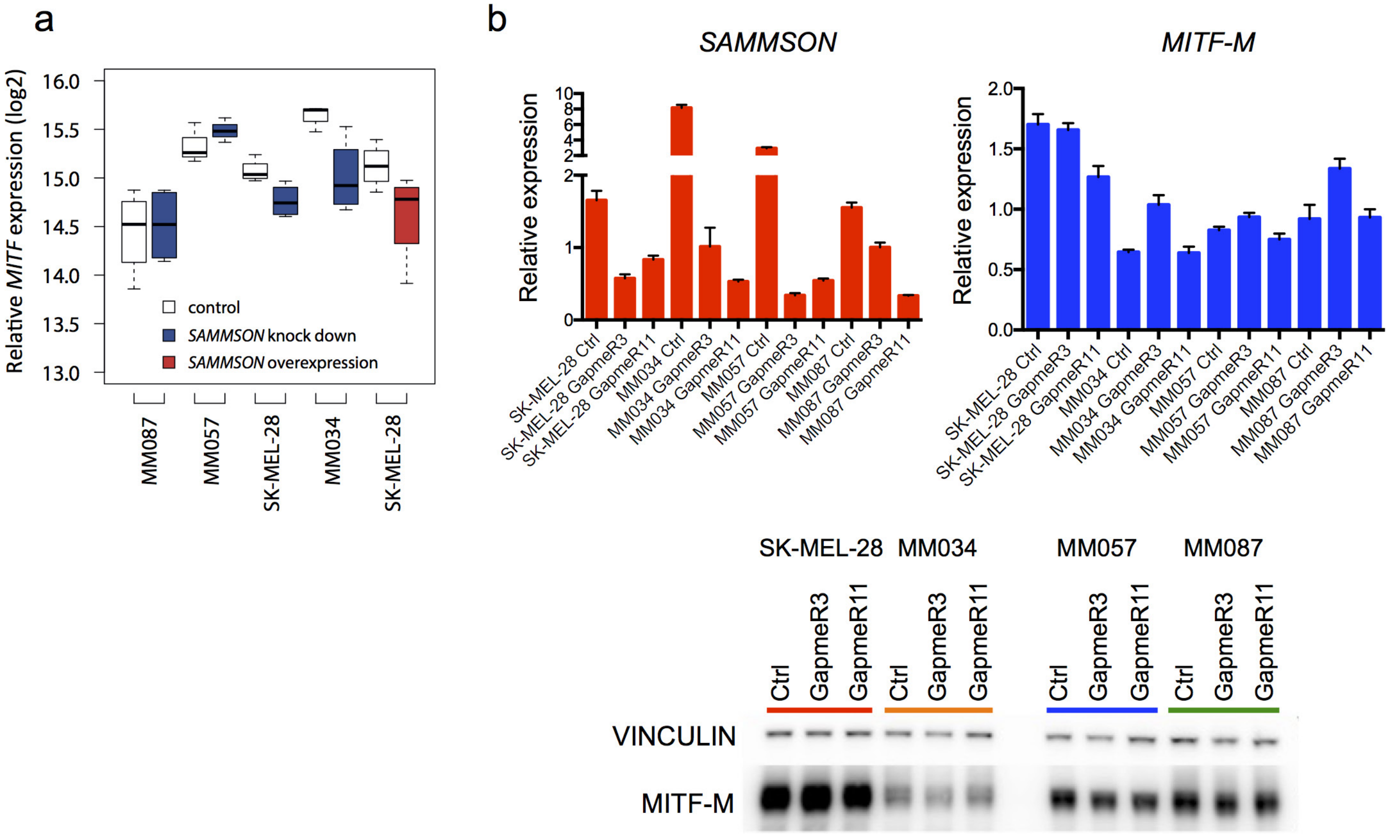
**Extended Data Figure 1 | *SAMMSON* is a lineage-specific lncRNA expressed in the vast majority of melanomas. a**, *SAMMSON* is a polyadenylated and multi-exonic lncRNA that contains one additional exon (in red) downstream of the four GENECODE19-annotated exons (blue). For each melanoma (SKCM) sample in the TCGA, the mapped RNA-seq data were converted into a coverage plot. The coverage data are normalized for library size and log1p transformed. Subsequently, the coverage data for all primary (SKCM.01) and metastatic samples (SKCM.06) in the region chromosome 3:70,040,000–70,140,000 is plotted as two heat maps. **b**, Cap analysis of gene expression-sequencing (CAGE-seq), RNA-seq data and RT–qPCR analyses from short-term melanoma cultures (MM lines) confirmed that *SAMMSON* is not a read-through transcript from the upstream *MITF* locus. CAGE-seq counts as defined by the FANTOM5 mammalian promoter expression atlas for two melanoma cell lines at the *SAMMSON* locus (panels 1 and 2), location of the a 3′ rapid amplification of cloned/cDNA ends (RACE) fragment for *SAMMSON* (panel 3) and RNA-seq counts from a primary melanoma tumour in the *SAMMSON* locus. **c**, *SAMMSON* and *MITF* copy number as measured by qPCR in short-term melanoma cultures and melanoma cell lines. Reference human genomic DNA was used as scaling control. Error bars represent s.d. of qPCR replicates ($n = 2$). A significant correlation between *MITF* and *SAMMSON* copy number was observed (bottom; Spearman's rank rho = 0.933, $P < 0.001$). **d**, Expression of *SAMMSON* in human short-term melanoma (MM) and NHME cultures relative to the expression average of three housekeepings (left) and correlation with MITF expression by western blot (right; for gel source data, see Supplementary Fig. 1). Error bars represent s.d. of three replicates ($n = 3$). **e**, Expression correlation between *MITF-M* and *SAMMSON* in melanoma clinical samples from the TCGA database. **f**, Read counts were generated from RNA-seq data from TCGA melanoma samples (SKCM) and normalized to the library size. Samples were subdivided into proliferative and invasive groups as described previously[3] and box plots were generated for *SAMMSON*, *ZEB1* and *MITF*. Differential expression analysis between the proliferative and invasive groups was done using edgeR50. lfc, log fold change; pval, uncorrected *P* value; fdr, false-discovery-rate-corrected pval. **g**, Relative expression in 60 cancer cell lines (NCI60 panel). **h**, Fraction of lncRNA expression variation ($n = 1,472$) across the NCI60 panel by cancer type according to a generalized additive model (GAM).



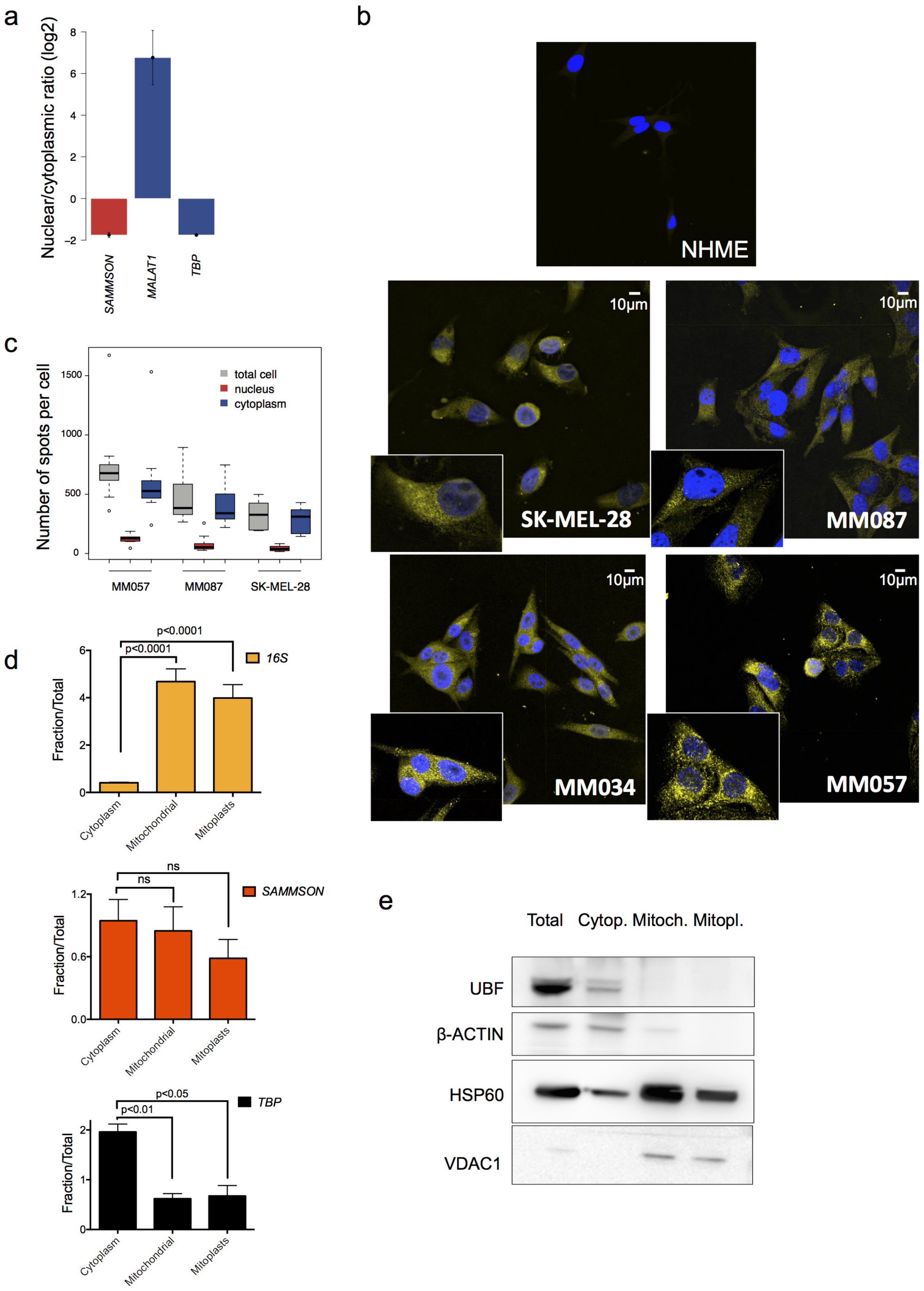
**Extended Data Figure 2 | *SAMMSON* expression in melanoma, but not other cancer, cell lines, is at least partly SOX10-dependent.** **a**, H3K27ac ChIP-seq data generated in house using a series of short-term melanoma cultures[2] were integrated with cancer cell lines data retrieved from ENCODE. A clear H3K27ac peak is present upstream *SAMMSON* in all, but one (MM001), melanoma lines. No peak is detected in the vast majority (19/20) of non-melanoma cancer cell lines, of which 9 are shown. **b**, UCSC screenshots of ChIP-seq data for SOX10, 3HA-MITF and H3K27ac at the *MITF* and *SAMMSON* loci in Mel501. **c**, ChIP-qPCR of endogenous SOX10 in 501Mel cells at the indicated loci. The IgG antibody was used as a negative control. SOX10 recruitment to its well-established targets *GPR110*, *TYR* and *SOX10* itself, but not to a non-SOX10 target *PRM1*, confirms the specificity of the SOX10 ChIP experiment. **c**, Western blotting analysis of total protein lysates of SK-MEL-28 confirming efficient knockdown of SOX10 and MITF. GAPDH was used as a loading control (for gel source data, see Supplementary Fig. 1). **e**, Fold change RNA expression levels in 501Mel cells transfected with a control siRNA pool (siCtrl) or pools targeting MITF (siMITF) or SOX10 (siSOX10).



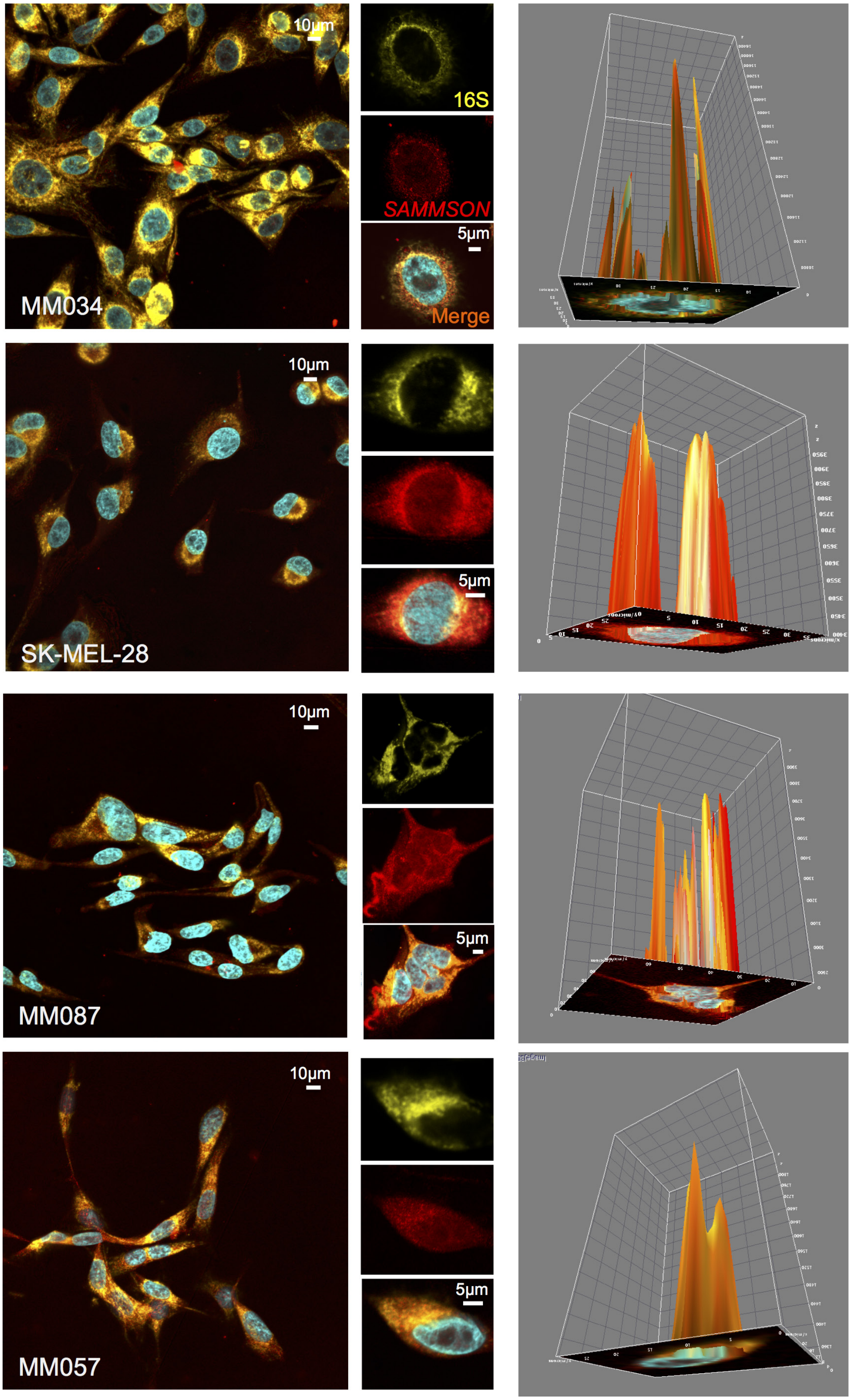
**Extended Data Figure 3 | *SAMMSON* promotes the *in vitro* growth and survival of human melanomas. a**, Colony formation assays 7 days after seeding of metastatic melanoma cells transfected with a GapmeR control (Ctrl), GapmeR3 and GapmeR11. **b**, Evaluation of cell death by co-staining for annexin V and propidium iodide (PI) followed by flow cytometric analysis. The graph is an average of three biological replicas and shows the percentage of cells alive, single positive or double positive. **c**, Efficiency of *SAMMSON* knockdown using an siRNA against *SAMMSON*. The expression of *SAMMSON* is relative to the average of three different housekeeping genes. **d**, Percentage of remaining living cells upon si*SAMMSON*, measured by flow cytometry, is indicated on *y*-axis ± s.e.m. **e**, Evaluation of the capacity of exogenus *SAMMSON* and *SAMMSON* mutants (*SAMMSON*gap3mut and *SAMMSON*gap11mut, in which mismatches were introduced into the GapmeR3 and GapmeR11 target sequences) to rescue cell death in SK-MEL-28 treated with GapmeRs. The percentage of remaining living cells is indicated on the *y*-axis ± s.e.m. **f**, Effect of a caspase-9 inhibitor on caspase-3/7 activation in SK-MEL-28 treated with GapmeRs. The graph is an average of three biological replicas; caspase-3/7 activity is relative to control sample (Ctrl) ± s.e.m. *P* value was calculated by ANOVA. **g**, Relative *SAMMSON* expression in MM001 cells transfected with an empty or *SAMMSON*-encoding expression vector of three different biological replicates ± s.e.m. *P* values were calculated by ANOVA. **h**, Colony formation assays 7 days after seeding 1,000, 5,000 or 10,000 MM001 as described in **g**. **i**, Colony formation assays 7 days after seeding of SK-MEL-28 transfected with a control GapmeR (Ctrl) or GapmeR3 and exposed to either vehicle or an $EC_{50}$ dose of vemurafenib (PLX4032) or pimasertib.



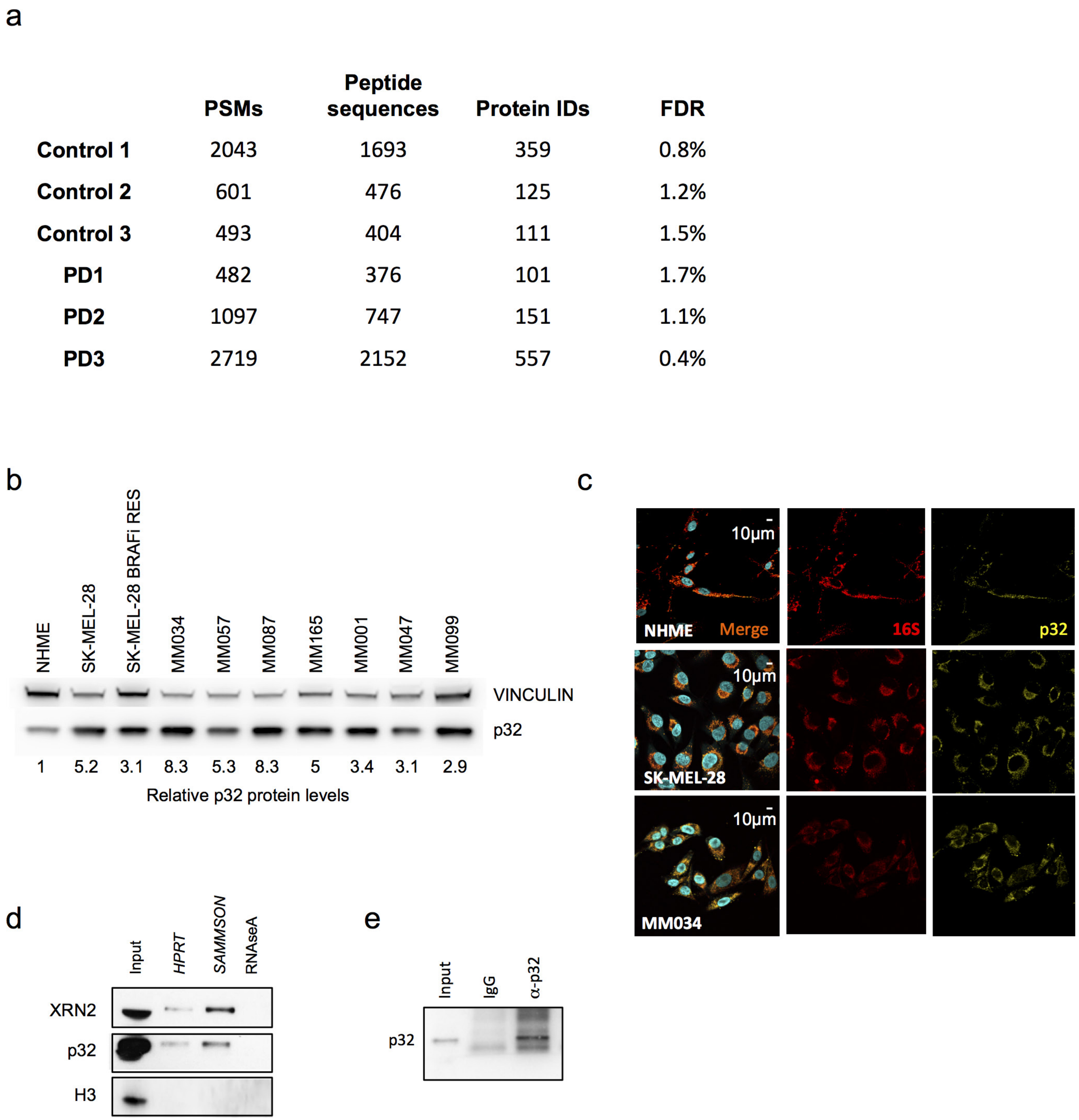
**Extended Data Figure 4 | *SAMMSON* does not regulate *MITF* expression in *cis*. a**, Relative expression of *MITF* as determined by microarray gene expression profiling in the indicated melanoma cell lines treated with GapmeR3 (*SAMMSON* knockdown) ($n > 3$) or expressing exogenous *SAMMSON* ($n = 4$); no significant differences in *MITF* expression were observed (limma, Benjamini–Hochberg adjusted $P > 0.05$), except in MM034, in which *SAMMSON* knockdown resulted in a 1.5-fold downregulation of *MITF* (limma, Benjamini–Hochberg adjusted $P = 0.013$). **b**, Validation of the array data in **a** by qPCR for *SAMMSON* (left) and *MITF-M* (right) in all the cell lines used for the arrays. Expression is relative to three different housekeeping genes. The graph shows an average of three different biological replicas ± s.e.m. The MITF-M protein levels were assessed by western blotting (bottom; for gel source data, see Supplementary Fig. 1).



**Extended Data Figure 5 | a, *SAMMSON* localizes primarily to the cytoplasm and largely co-localizes with mitochondria.** Quantification of *SAMMSON* in nuclei and cytoplasm of SK-MEL-28. Data are expressed as nuclear/cytoplasmic ratio ± s.e.m. Data are shown for *MALAT1* (exclusively nuclear) and *TBP* (cytoplasmic). The graph shows an average of three different fractionation experiments. **b**, *SAMMSON* RNA-FISH in a panel of melanoma cell lines and NHMEs. *SAMMSON* is shown in yellow and DAPI in blue. **c**, Quantification of *SAMMSON* RNA-FISH results described in **b**. Number of fluorescent spots in total per cell, nucleus and cytoplasm of MM057 ($n = 10$), MM087 ($n = 10$) and SK-MEL-28 ($n = 7$) melanoma cells are shown. **d**, Quantification of *SAMMSON* in cytoplasm, mitochondria and mitoplasts of SK-MEL-28. Data are expressed as fraction/total ratio ± s.e.m. Data are shown for mitochondrial 16S rRNA (exclusively mitochondrial) and *TBP* (cytoplasmic). The graph is an average of three different fractionation experiments. **e**, The purity of the fractions was assessed by western blotting using nuclear (UBF1), cytoplasmic (β-actin) and mitochondrial markers (HSP60 and VDAC1; for gel source data, see Supplementary Fig. 1).

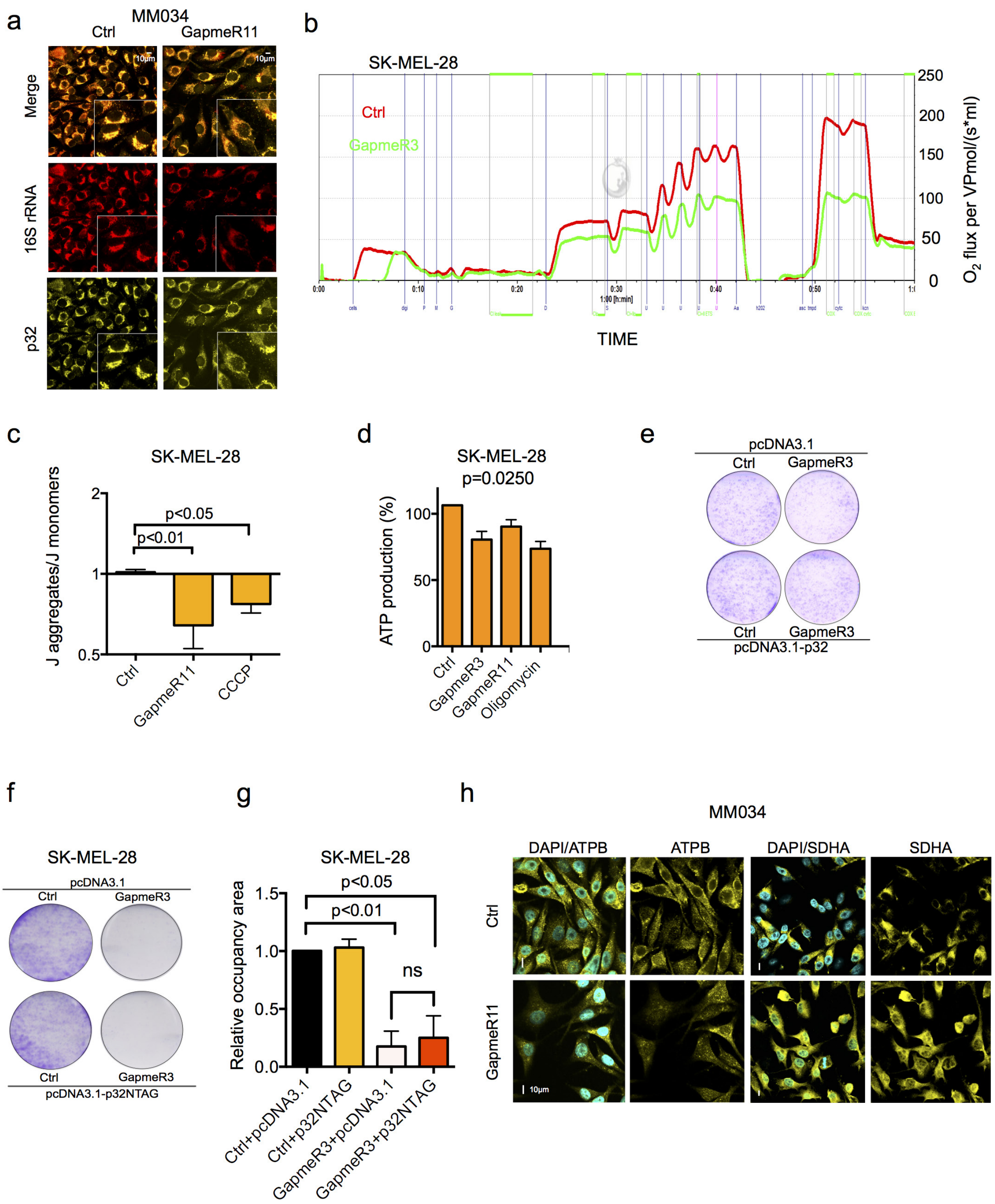


**Extended Data Figure 6 | A large fraction of cytoplasmic *SAMMSON* co-localizes with mitochondria.** *SAMMSON* and mitochondrial 16S rRNA RNA-FISH in four different melanoma cell lines. *SAMMSON* probes, labelled with Quasar570, are shown in red and 16S rRNA probes, labelled with Quasar670, are shown in yellow; DAPI is in cyan.

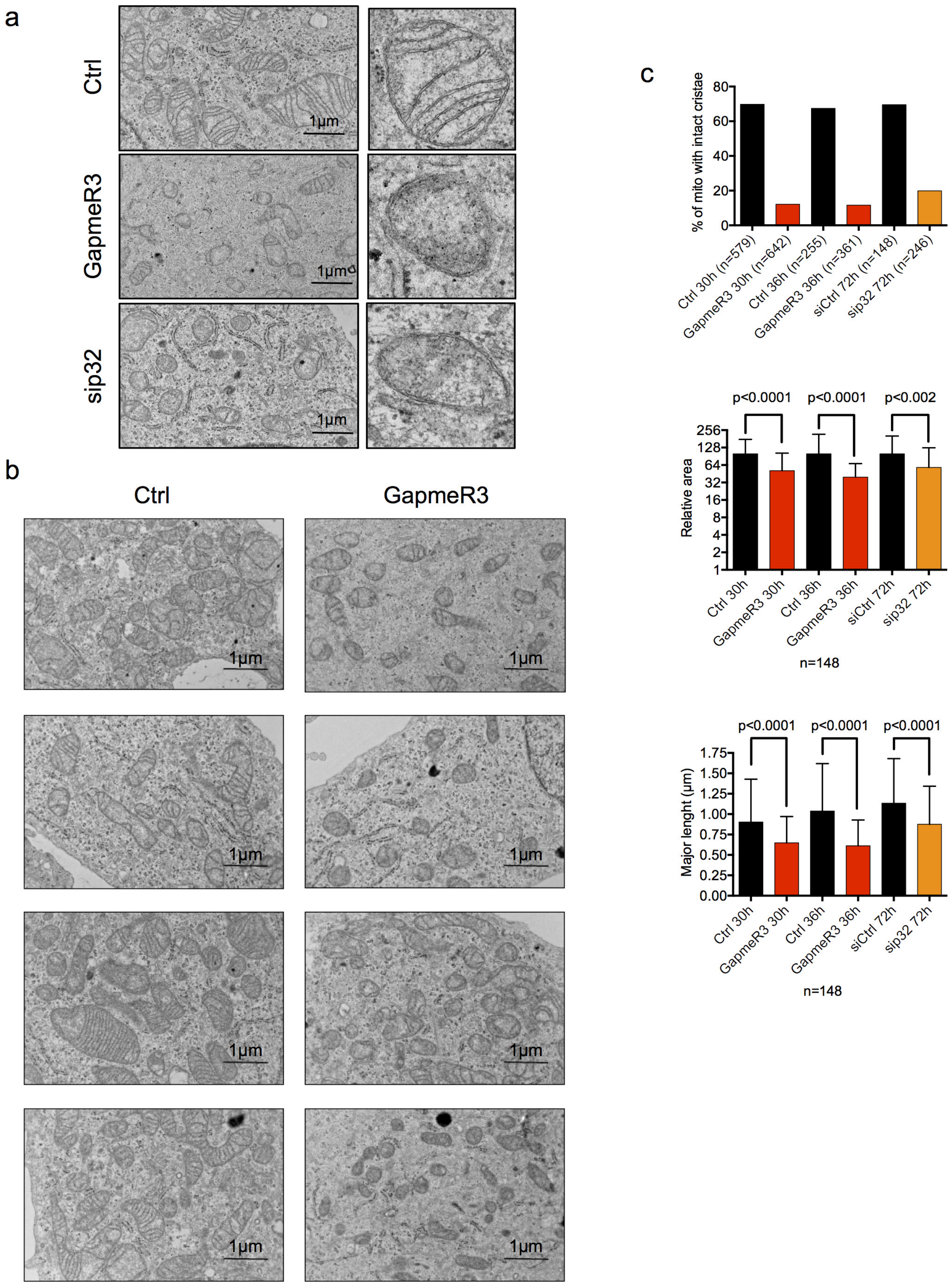


|  | PSMs | Peptide sequences | Protein IDs | FDR |
|---|---|---|---|---|
| Control 1 | 2043 | 1693 | 359 | 0.8% |
| Control 2 | 601 | 476 | 125 | 1.2% |
| Control 3 | 493 | 404 | 111 | 1.5% |
| PD1 | 482 | 376 | 101 | 1.7% |
| PD2 | 1097 | 747 | 151 | 1.1% |
| PD3 | 2719 | 2152 | 557 | 0.4% |

**Extended Data Figure 7 | RAP–MS identifies the mitochondrial protein p32 as a *SAMMSON* interactor. a**, The 'Metrics' table provides an overview of the MS experiment. For every MS analysis, the table contains the number of analysed spectra, the total number of identified spectra (peptide to spectra matches (PSMs)), the number of distinct peptide sequences, protein numbers and a false discovery rate (FDR) estimation based on searches against a reversed database. **b**, Western blot for p32 in NHMEs and in a panel of short-term melanoma cultures. Vinculin is used as a loading control and normalizer for the quantification. **c**, RNA-FISH for 16S rRNA (in red) and immunofluorescence for p32 (in yellow) in NHMEs and in a panel of melanoma cell lines. DAPI is in cyan. **d**, Pulldown of *SAMMSON* (and *HRPT)* under native conditions and upon incubation with RNase A, followed by western blotting. **e**, Western blot confirming enrichment of p32 following immunoprecipitation with anti-p32 antibodies. **b**, **d**, **e**, For gel source data, see Supplementary Fig. 1.



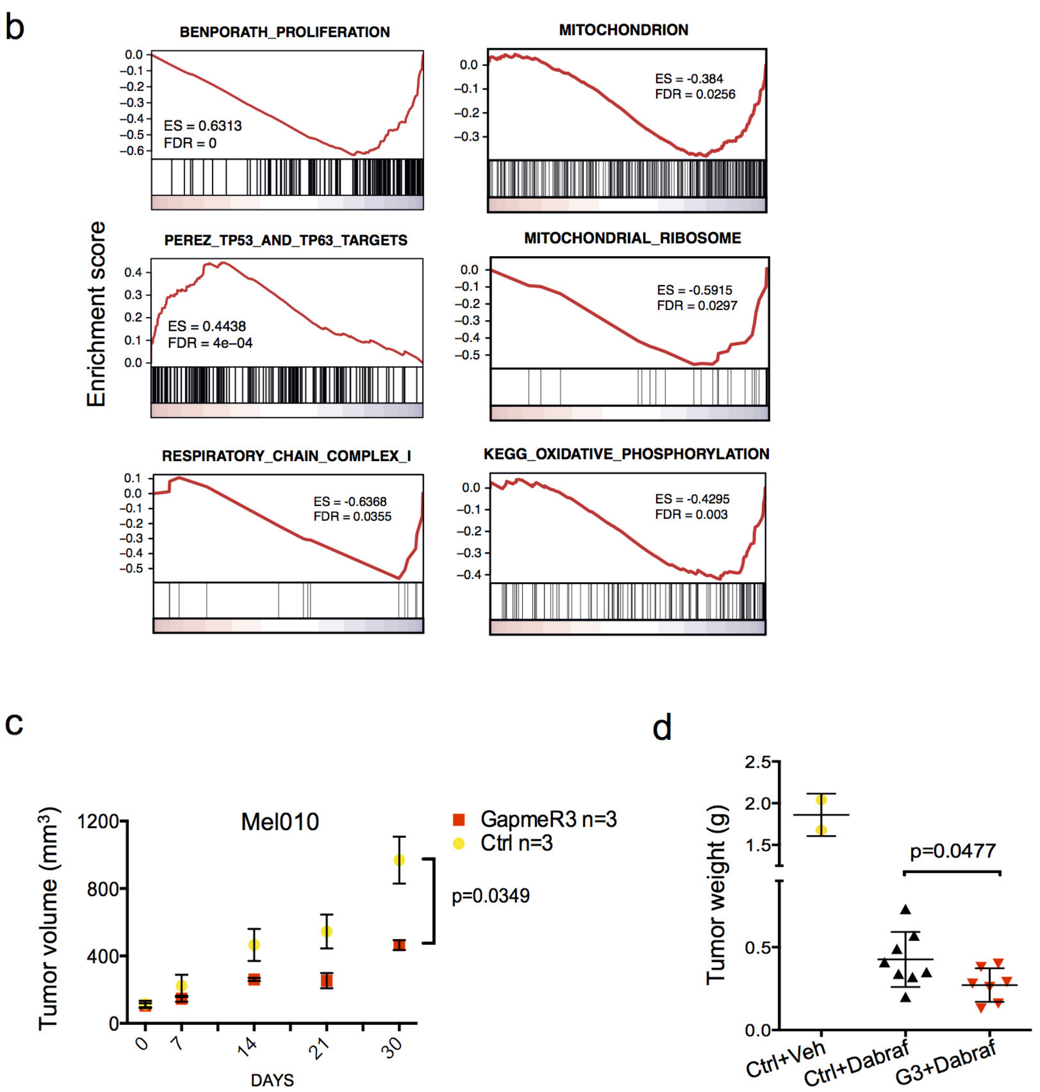
**Extended Data Figure 8 | *SAMMSON* modulates mitochondrial metabolism in a p32-dependent manner. a**, RNA-FISH for 16S rRNA (in red) and immunofluorescence for p32 (in yellow) in melanoma cells (MM034) transfected with a control GapmeR (Ctrl) or with GapmeR11. Magnification, ×200. **b**, Evaluation of OXPHOS complexes functionality by high-resolution respirometry in SK-MEL-28 treated with GapmeR3. The graph shows one representative experiment. **c**, JC-1 reveals a decrease in mitochondrial membrane potential upon GapmeR11 transfection, as assessed by the ratio between J-aggregates and J-monomers. The graph is an average of four different biological replicates ± s.e.m. *P* values were calculated by ANOVA. **d**, Evaluation of ATP production in SK-MEL-28 transfected with GapmeRs and exposed to oligomycin (used here as control). ATP production measured by luciferase is expressed as a percentage, an average of three different experiments, relative to the control sample (Ctrl) ± s.e.m. *P* values were calculated by ANOVA. **e**, Colony formation assays 5 days after seeding SK-MEL-28 transfected with a control GapmeR (Ctrl) or GapmeR3, and either with an empty or a p32-expressing vector. **f**, Colony formation assays 5 days after seeding, showing cell growth of SK-MEL-28 transfected with a control GapmeR (Ctrl) or GapmeR3, and either with an empty vector or a vector expressing a tagged version of p32 that cannot localize to the mitochondria. **g**, Quantification of the colony assay described in **f**. The data represent the density (occupancy area) relative to the Ctrl + pcDNA3.1 sample. The data are presented as average of three different biological replicates ± s.e.m. *P* values were calculated by ANOVA. **h**, Immunofluorescence using antibodies directed against SDHA and ATPB in MM034 melanoma cells upon GapmeR11 transfection. Magnification, ×600.



**Extended Data Figure 9 | *SAMMSON* and p32 silencing affects mitochondria integrity. a**, **b**, Representative electron microscopy images of melanoma cells transfected with Ctrl GapmeR and GapmeR3 or with siRNA targeting p32. **c**, Quantification of percentage of mitochondria with intact cristae (top) and area and length (middle and bottom) in cells described in **b**; the total number of mitochondria evaluated per condition is indicated on the *x*-axis.

a

| Patient | Mel006 | Mel010 |
|---|---|---|
| Gender | Female | Female |
| Year of birth | 1947 | 1955 |
| Biopsy origin | In transit metastasis | Lymph node |
| Biopsy site | Arm | Chest |
| Treatment prior to biopsy | none | none |
| BRAF V600 status | V600E | V600E |
| NRAS Q61 status | WT | WT |
| p53 status | WT (SNP c.215C>G; p.P72R) | WT (SNP c.215C>G;p.P72R) |



**Extended Data Figure 10 | *SAMMSON* silencing decreases melanoma growth *in vivo*. a**, Table describes the origin and *BRAF*, *NRAS* and *TP53* mutational status of the melanoma lesions that were used to generate the Mel006 and Mel010 PDX models. **b**, Gene set enrichment analyses among differentially expressed genes in melanoma lesions obtained from Mel006 PDX mice treated (i.v. injections) with a control GapmeR (Ctrl) or GapmeR3. **c**, Tumour volume of cohorts of Mel010 PDX mice treated (intra-tumour injections) with a control GapmeR (Ctrl) or GapmeR3. Data are means ± s.d. of three different biological replicates (*P* value was calculated by two-ways ANOVA). **d**, Tumour weight of the melanoma lesions derived from PDX mice (Mel006) treated with combinations of control GapmeR (Ctrl) and GapmeR3 with either vehicle or dabrafenib by daily oral gavage (vehicle or dabrafenib) and i.v. injection of the GapmeRs every 2 days. *P* value was calculated by *t*-test.

doi:10.1038/nature17186

# The amino acid sensor GCN2 controls gut inflammation by inhibiting inflammasome activation

Rajesh Ravindran[1]*, Jens Loebbermann[1]*, Helder I. Nakaya[2], Nooruddin Khan[3], Hualing Ma[1], Leonardo Gama[2], Deepa K. Machiah[4], Benton Lawson[5], Paul Hakimpour[1], Yi-chong Wang[1], Shuzhao Li[1], Prachi Sharma[4], Randal J. Kaufman[6], Jennifer Martinez[7] & Bali Pulendran[1]

**The integrated stress response (ISR) is a homeostatic mechanism by which eukaryotic cells sense and respond to stress-inducing signals, such as amino acid starvation. General controlled non-repressed (GCN2) kinase is a key orchestrator of the ISR, and modulates protein synthesis in response to amino acid starvation. Here we demonstrate in mice that GCN2 controls intestinal inflammation by suppressing inflammasome activation. Enhanced activation of ISR was observed in intestinal antigen presenting cells (APCs) and epithelial cells during amino acid starvation, or intestinal inflammation. Genetic deletion of *Gcn2* (also known as *Eif2ka4*) in CD11c$^+$ APCs or intestinal epithelial cells resulted in enhanced intestinal inflammation and T helper 17 cell (T$_H$17) responses, owing to enhanced inflammasome activation and interleukin (IL)-1β production. This was caused by reduced autophagy in *Gcn2*$^{-/-}$ intestinal APCs and epithelial cells, leading to increased reactive oxygen species (ROS), a potent activator of inflammasomes[1]. Thus, conditional ablation of *Atg5* or *Atg7* in intestinal APCs resulted in enhanced ROS and T$_H$17 responses. Furthermore, *in vivo* blockade of ROS and IL-1β resulted in inhibition of T$_H$17 responses and reduced inflammation in *Gcn2*$^{-/-}$ mice. Importantly, acute amino acid starvation suppressed intestinal inflammation via a mechanism dependent on GCN2. These results reveal a mechanism that couples amino acid sensing with control of intestinal inflammation via GCN2.**

The immune system can sense pathogens through pathogen recognition receptors[2], but emerging evidence suggests that it can also sense and respond to environmental changes that cause cellular stress[3]. The ISR is an evolutionarily ancient mechanism that enables eukaryotic cells to sense and respond to diverse stress signals, such as amino acid starvation and endoplasmic reticulum stress[4]. The four known sensors of the ISR include: GCN2, protein kinase R (PKR), haem-regulated inhibitor (HRI) and PKR-like endoplasmic reticulum kinase (PERK)[4]. GCN2 senses amino acid depletion, PERK senses endoplasmic reticulum stress, and PKR can recognize viral double-stranded RNA[4]. Activation of HRI is induced by haem deficiency[5], and is important for the survival of erythroid precursors. Activation of each of these four sensors results in phosphorylation of eukaryotic initiation factor 2α (eIF2α), leading to the initiation of global translational arrest[4]. Recent evidence suggests a crosstalk between the ISR and the immune system[3]. Thus, our recent systems-based analysis of immune responses to the yellow fever vaccine (YF-17D) in humans revealed a correlation between the expression of GCN2 in the blood and the magnitude of the later CD8$^+$ T-cell response[6]. Furthermore, YF-17D induced GCN2 activation in dendritic cells, resulting in enhanced autophagy and antigen presentation[7]. Whether GCN2 can modulate immune responses during conditions of amino acid restriction remains unexplored. This is particularly relevant in the intestine, where the immune system has to endure dynamic changes in nutrient bioavailability. We thus determined whether GCN2 impacts immune homeostasis in the intestine.

Phosphorylated eIF2α was detected in intestinal dendritic cells, macrophages and epithelial cells under steady-state and inflammatory conditions (Extended Data Fig. 1a). Furthermore, expression of phosphorylated PKR, PERK, eIF2α and GCN2 could be detected in tissues from healthy and inflamed human colon (Extended Data Fig. 1b). Analysis of public gene expression databases revealed that the expression of genes encoding GCN2 and other eIF2α kinases was highest in the colon, relative to other organs (Extended Data Fig. 1c). Interestingly, there was a higher expression of genes encoding GCN2, PERK and PKR in ulcerative colitis and Crohn's disease, relative to healthy controls[8,9] (Extended Data Fig. 1d).

To investigate the functions of GCN2 *in vivo*, we analysed the structure and morphology of gut tissue isolated from the *Gcn2*$^{-/-}$ mice. Ki-67 and chromogranin A staining in small and large intestines were unaffected in *Gcn2*$^{-/-}$ mice, suggesting that GCN2 is not required for steady-state cell differentiation and proliferation in the intestine (Extended Data Fig. 2a, b, d). *Gcn2*$^{-/-}$ mice had normal Paneth cell granules, as evident with lysozyme staining (Extended Data Fig. 2c), and did not exhibit any spontaneous gut inflammation up to 45 weeks of age.

We then assessed the impact of GCN2 deficiency on acute colitis by challenging the mice with 2% dextran sodium sulfate (DSS), a chemical irritant that induces inflammation with the clinical and histological features of inflammatory bowel disease in mice[10]. Upon DSS administration, *Gcn2*$^{-/-}$ mice exhibited enhanced severity of colitis compared with littermates, including greater weight loss, inflammation, T$_H$17 responses and colon shortening (Fig. 1a–c and Extended Data Fig. 3a–c). Histopathological analysis revealed severe mucosal epithelial erosion, displacement and crypt loss (Extended Data Fig. 3a). Consistent with enhanced gut inflammation, we observed a severely impaired epithelial barrier, evidenced by increased intestinal permeability (Extended Data Fig. 3d). These differences were not due to differences in the expression of antimicrobial defensins between wild-type and *Gcn2*$^{-/-}$ mice (Extended Data Fig. 3e).

To assess potential roles for APCs versus epithelial cells in mediating the effects of GCN2, we generated mice lacking GCN2 specifically in epithelial cells (*Gcn2*$^{fl/fl}$ villin *cre*$^+$; referred to as *Gcn2*$^{\Delta\text{villin}}$ hereafter) (Fig. 1d–f and Extended Data Fig. 3a–c), or in CD11c$^+$ APCs (*Gcn2*$^{fl/fl}$ *Cd11c cre*$^+$; referred to as *Gcn2*$^{\Delta\text{APC}}$ hereafter) (Fig. 1g–i and Extended Data Fig. 3a–c). DSS induced enhanced colitis in both strains,

[1]Emory Vaccine Center, Yerkes National Primate Research Center, 954 Gatewood Road, Atlanta, Georgia 30329, USA. [2]School of Pharmaceutical Sciences, University of São Paulo, São Paulo 05508, Brazil. [3]Department of Biotechnology and Bioinformatics, School of Life Sciences, University of Hyderabad, Hyderabad 500 046, India. [4]Division of Pathology, Yerkes National Primate Research Center, 954 Gatewood Road, Atlanta, Georgia 30329, USA. [5]Virology Core, Emory Vaccine Center and Yerkes National Primate Research Center, 954 Gatewood Road, Atlanta, Georgia 30329, USA. [6]Degenerative Disease Program, Sanford Burnham Prebys Medical Discovery Institute, 10901 North Torrey Pines Road, La Jolla, California 92037 USA. [7]National Institute of Environmental Health Sciences, Mail Drop D2-01 Research Triangle Park, North Carolina 27709, USA.
*These authors contributed equally to this work.

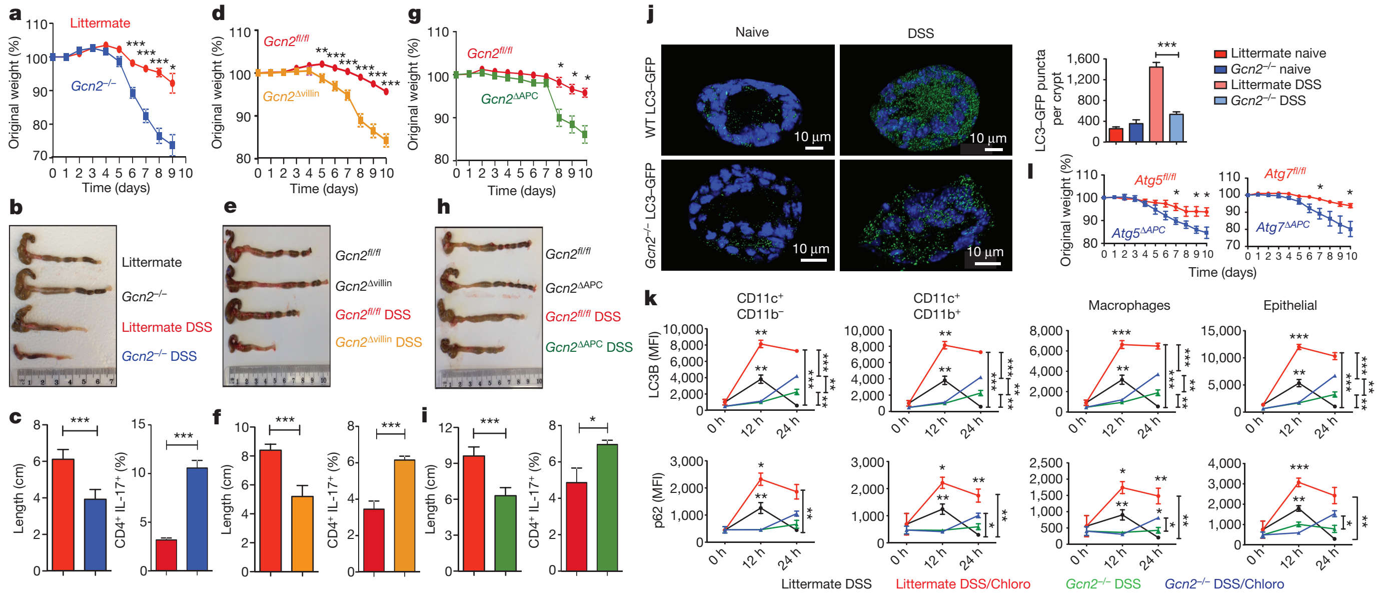


**Figure 1 | GCN2 activation in APCs and epithelial cells suppresses intestinal inflammation by a mechanism dependent on autophagy.** **a**–**c**, GCN2 deficiency leads to loss of body weight, colon shortening and enhanced production of IL-17 by colonic CD4$^+$ T cells. **d**–**i**, Expression of GCN2 in epithelial cells (*Gcn2*$^{\Delta\text{villin}}$) (**d**–**f**) or APCs (*Gcn2*$^{\Delta\text{APC}}$) (**g**–**i**) protects mice from DSS-induced colitis. **j**, LC3–GFP expression and the GFP puncta counts in the crypts (three-dimensional) from *Gcn2*$^{-/-}$ LC3–GFP and LC3–GFP mice before and 12 h after DSS. WT, wild type. **k**, Mean fluorescence intensity (MFI) comparison of LC3B and p62 expression with and without chloroquine (Chloro) on APC subsets and epithelial cells by flow cytometry before and after 3% DSS administration. **l**, Comparison of body weights of *Atg5*$^{\Delta\text{APC}}$ and *Atg7*$^{\Delta\text{APC}}$ mice to littermate controls subjected to acute 2% DSS-induced colitis. Data are representative of three separate experiments ($n = 5$). $^*P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0005$. Error bars indicate mean ± standard error of the mean (s.e.m.), two-tailed unpaired Student's *t*-test.

evidenced by weight loss, colon shortening and increased T$_H$17 responses (a surrogate readout of intestinal inflammation) relative to littermate controls (Fig. 1 and Extended Data Fig. 3b, c). Consistent with a role for GCN2 in APCs, isolated intestinal dendritic cells from *Gcn2*$^{-/-}$ mice could stimulate enhanced IL-17 production from antigen-specific CD4$^+$ T cells, *in vitro* (Extended Data Fig. 3f). Collectively, these findings demonstrate that GCN2 deficiency in both epithelial cells and APCs results in enhanced inflammation and DSS-induced colitis.

Since PERK activation by endoplasmic reticulum stress is also known to be an important component of the host ISR[11], we generated mice lacking PERK in epithelial cells (*Perk*$^{fl/fl}$ villin *cre*$^+$ (*Perk* also known as *Eif2ak3*); referred to as *Perk*$^{\Delta\text{villin}}$ hereafter) or APCs (*Perk*$^{fl/fl}$ *Cd11c cre*$^+$; referred to as *Perk*$^{\Delta\text{APC}}$ hereafter) to study the role of PERK in intestinal inflammation (Extended Data Fig. 4). PERK-deficient strains were challenged with 2% DSS, and their symptoms and pathology were compared with littermate controls. Both *Perk*$^{\Delta\text{villin}}$ and *Perk*$^{\Delta\text{APC}}$ strains exhibited little or no differences relative to littermates in intestinal inflammation induced by DSS (Extended Data Fig. 4).

As ISR kinases exert their function by phosphorylating Ser51 on eIF2α, we assessed the impact of eIF2α on intestinal inflammation. We generated mice conditionally lacking Ser51 eIF2α phosphorylation in epithelial cells (*Eif2a*$^{fl/fl}$ villin *cre*$^+$; referred to as *Eif2a*$^{\Delta\text{villin}}$ hereafter) as previously described[12,13] and APCs (*Eif2a*$^{fl/fl}$ *Cd11c cre*$^+$; referred to as *Eif2a*$^{\Delta\text{APC}}$ hereafter) (Extended Data Fig. 5). *Eif2a*$^{\Delta\text{villin}}$ mice exhibited enhanced weight loss and elevated T$_H$17 response relative to littermate controls, and *Eif2a*$^{\Delta\text{APC}}$ mice exhibited enhanced T$_H$17 responses (Extended Data Fig. 5), consistent with a recent report on a role for eIF2α in mediating protection against gut inflammation[13]. However, these effects were more modest than those observed in *Gcn2*$^{-/-}$ mice (Fig. 1), suggesting additional eIF2α-independent mechanisms.

Recent studies indicate a role for GCN2 in promoting autophagy[7,14]. Given its importance in regulating inflammation at mucosal sites[15], we hypothesized that defective autophagy may mediate enhanced gut inflammation in *Gcn2*$^{-/-}$ mice. We analysed expression of the autophagy protein LC3 using a knock-in reporter strain[16], and observed high LC3 expression in colonic APCs and epithelial cells, which was indicative of constitutive autophagy (Extended Data Fig. 6a). To study the role of GCN2 in mediating autophagy at mucosal sites, we generated GCN2-deficient autophagic reporter mice (*Gcn2*$^{-/-}$ × LC3–GFP), and additionally examined expression of p62, another marker for autophagy. In naive mice, expression of both LC3 and p62 were similar in the intestinal APCs and epithelial cells of *Gcn2*$^{-/-}$ mice and littermates (Fig. 1j, k). However, we observed a significant increase in the number of LC3–GFP puncta in the crypts of wild-type mice compared with the *Gcn2*$^{-/-}$ mice after oral administration of DSS (Fig. 1j). Additionally, we observed that intestinal APCs and epithelial cells from *Gcn2*$^{-/-}$ mice have lower levels of LC3B and p62 relative to cells from wild-type mice (Fig. 1k and Extended Data Fig. 6b). To determine whether the observed reduction in autophagy in *Gcn2*$^{-/-}$ mice was due to impaired induction or enhanced degradation of autophagosomes, we assessed LC3B and p62 levels in intestinal APCs and epithelial cells from wild-type or *Gcn2*$^{-/-}$ mice with or without chloroquine, an inhibitor of autophagy degradation (Fig. 1k and Extended Data Fig. 6b). Blocking the degradation of autophagosomes with chloroquine in wild-type mice resulted in greater accumulation of LC3B and p62 at 12 h after DSS (Fig. 1k and Extended Data Fig. 6b). Importantly, the accumulated form (LC3B and p62) was significantly lower in *Gcn2*$^{-/-}$ cells, indicating reduced autophagy flux relative to wild-type mice (Fig. 1k and Extended Data Fig. 6b). Similar results were observed at 24 h, although chloroquine-treated *Gcn2*$^{-/-}$ mice displayed enhanced accumulation of LC3B and p62, albeit at lower levels than wild-type mice (Fig. 1k and Extended Data Fig. 6b). The specificity of the LC3B antibody to LC3BII was confirmed using digitonin to retain specifically the membrane-bound LC3II within cells[17] (Extended Data Fig. 6c, d).

These findings suggested that defective autophagy in intestinal cells may mediate enhanced inflammation in *Gcn2*$^{-/-}$ mice.

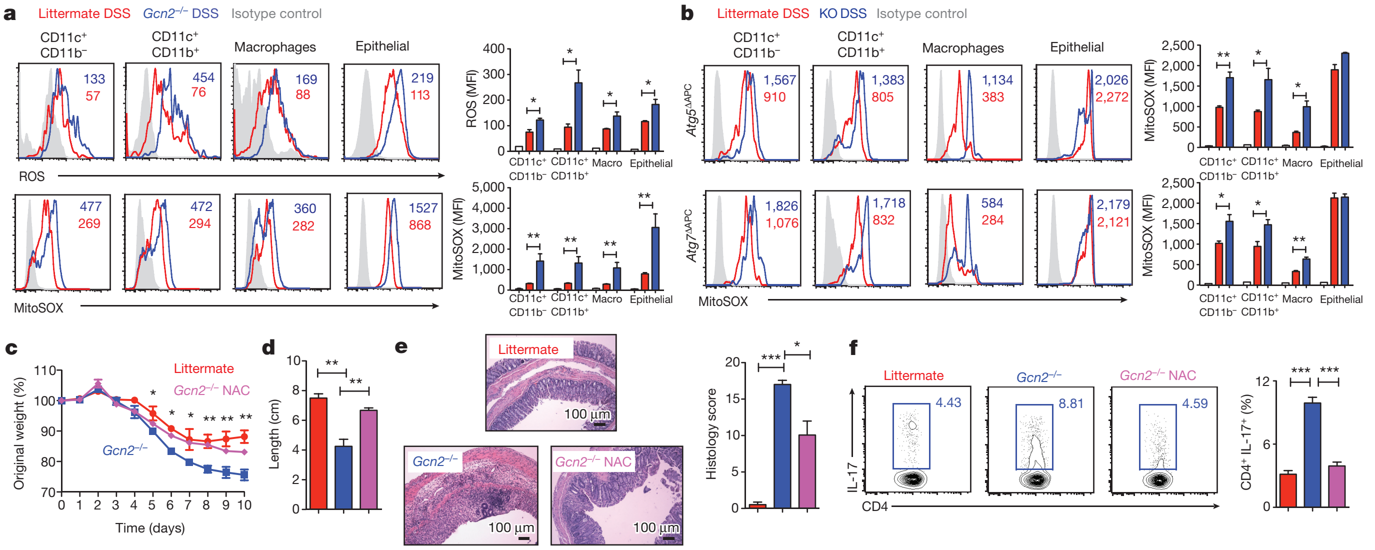


**Figure 2 | GCN2 suppresses ROS activity and intestinal inflammation.** **a**, Histograms and MFI quantification of ROS and mitochondrial ROS in colonic APC subsets and epithelial cells isolated from wild-type and *Gcn2*$^{-/-}$ mice after 5 days of DSS. Macro, macrophages. **b**, *Atg5*$^{\Delta\text{APC}}$ and *Atg7*$^{\Delta\text{APC}}$ mice compared with littermates before and after 5 days of DSS. KO, knockout. **c–f**, NAC protects *Gcn2*$^{-/-}$ mice from weight loss (**c**), reduces colon length shortening (**d**), ameliorates decreased pathology (**e**) and reduces colonic T$_{\text{H}}$17 frequencies (**f**) induced by DSS. Data are representative of three experiments ($n = 5$). *$P < 0.05$; **$P < 0.005$, ***$P < 0.0005$. Error bars indicate mean ± s.e.m.

We therefore generated *Atg5*$^{fl/fl}$ *Cd11c cre* (Atg5$^{\Delta APC}$) and *Atg7*$^{fl/fl}$ *Cd11c cre* (Atg7$^{\Delta APC}$) mice that are conditionally deficient in the autophagy proteins Atg5 and Atg7 in CD11c$^{+}$ APCs (Extended Data Fig. 7). Upon treatment with 2% DSS, both the Atg5$^{\Delta APC}$ and Atg7$^{\Delta APC}$ strains exhibited greater weight loss (Fig. 1l), enhanced shortening of colon length, T$_{\text{H}}$17 responses and immunopathology (Extended Data Fig. 7) compared with littermate controls, indicating a role for APC-intrinsic autophagy in regulating inflammation in *Gcn2*$^{-/-}$ mice.

Autophagy can limit ROS abundance during colitis[15], and impaired autophagy results in abnormal mitochondrial function and oxidative stress[15], which is a characteristic feature of inflammatory bowel disease[18]. Hence we studied the role of ROS in mediating DSS-induced inflammation using cell-permeant 2′,7′-dichlorodihydrofluorescein diacetate (H2DCFDA), a fluorescent probe that reacts with numerous types of ROS[19]. After oxidation by ROS, the non-fluorescent H2DCFDA is converted to fluorescent 2′,7′-dichlorofluorescein (DCF)[19], which was detected in intestinal APCs and epithelial cells (Fig. 2a and Extended Data Fig. 8a, b) by flow cytometry. *Gcn2*$^{-/-}$ mice exhibited significantly higher levels of ROS compared with littermate controls, indicating enhanced oxidative stress in the colon (Fig. 2a) and small intestine (Extended Data Fig. 8a, b). We also analysed the levels of mitochondrial ROS in the large and small intestine using MitoSOX, a fluorogenic dye that specifically detects mitochondrial superoxide[20]. *Gcn2*$^{-/-}$ mucosal cells produced excess mitochondrial ROS in comparison with the littermate controls in the colon (Fig. 2a) and small intestine (Extended Data Fig. 8a, b). To determine whether autophagy regulated mitochondrial ROS, we analysed superoxide levels in colonic cells isolated from Atg5$^{\Delta APC}$ and Atg7$^{\Delta APC}$ mice after DSS treatment. As in the *Gcn2*$^{-/-}$ strain, there was higher production of mitochondrial ROS in Atg5$^{\Delta APC}$ and Atg7$^{\Delta APC}$ mice compared with littermate controls (Fig. 2b). Furthermore, blockade of ROS via administration of the antioxidant *N*-acetyl-L-cysteine (NAC) *in vivo* led to reduced disease severity, and reduction of T$_{\text{H}}$17 responses in *Gcn2*$^{-/-}$ mice (Fig. 2c–f). Thus, these data demonstrate a key role for ROS in mediating the enhanced inflammation observed in *Gcn2*$^{-/-}$.

Oxidative mitochondrial stress is known to be involved in the activation of the inflammasome pathway[1,15]. We therefore hypothesized that excess ROS enhanced inflammasome activation in the *Gcn2*$^{-/-}$ cells under inflammatory conditions. *Gcn2*$^{-/-}$ dendritic cells produced excess amounts of cleaved IL-1β and cleaved caspases when subjected to amino acid starvation (Fig. 3a). Additionally, there was higher production of pro-IL-1β in the colonic macrophages and dendritic cells isolated from DSS-treated *Gcn2*$^{-/-}$ mice in the large (Fig. 3b) and small intestine (Extended Data Fig. 8b). *In vivo* blockade of IL-1β with a neutralizing antibody in *Gcn2*$^{-/-}$ mice ameliorated the deleterious effects of DSS (Fig. 3c, d and Extended Data Fig. 8c), and significantly reduced intestinal T$_{\text{H}}$17 response (Fig. 3e). However, there were no detectable effects on histopathology, possibly due to incomplete neutralization of IL-1β (Extended Data Fig. 8c). Additionally, we observed that increased inflammation and T$_{\text{H}}$17 responses in the *Gcn2*$^{-/-}$ mice were negated by the deletion of the inflammasome adaptor protein apoptosis-associated speck-like protein containing a CARD (ASC) (Fig. 3f–i), demonstrating a clear role for inflammasome activation in mediating the enhanced inflammation in *Gcn2*$^{-/-}$ mice.

Given the importance of GCN2 in sensing amino acid starvation, we hypothesized that mice fed an amino-acid-deprived diet might display enhanced activation of ISR in intestinal cells, resulting in dampened inflammation. Therefore we fed mice a reduced amino acid diet (2% protein/weight versus 16% in control mice) and observed rapid activation of phosphorylated (p)-eIF2α on intestinal APCs and epithelial cells (Fig. 4a). Intestinal cells isolated from wild-type mice that were on a low protein diet rapidly upregulated autophagy in comparison to those from *Gcn2*$^{-/-}$ mice (Fig. 4b). Consistent with this, mass spectrometric analysis of free cytosolic amino acids revealed reduced levels of specific amino acids in colonic epithelial cells and APCs isolated from mice on a 2% low protein diet, relative to the corresponding cell types from control mice, as well as in mice on DSS (data not shown).

Interestingly, diets lacking in selective amino acids can also preferentially activate the GCN2 pathway[21–23]. We next asked whether lowering of proteins (2% protein) or selective depletion of individual essential amino acids such as leucine (Leu$^{-}$) impacted intestinal inflammation (Fig. 4). Thus wild-type or *Gcn2*$^{-/-}$ mice were fed amino-acid-restricted or control diets (16% protein) and subsequently challenged with 3% DSS in their drinking water (Fig. 4c–e). Three per cent DSS was administered to induce enhanced inflammation in wild-type mice, so as to be able to reveal the putative protective effects of amino acid starvation on inflammation. Mice on

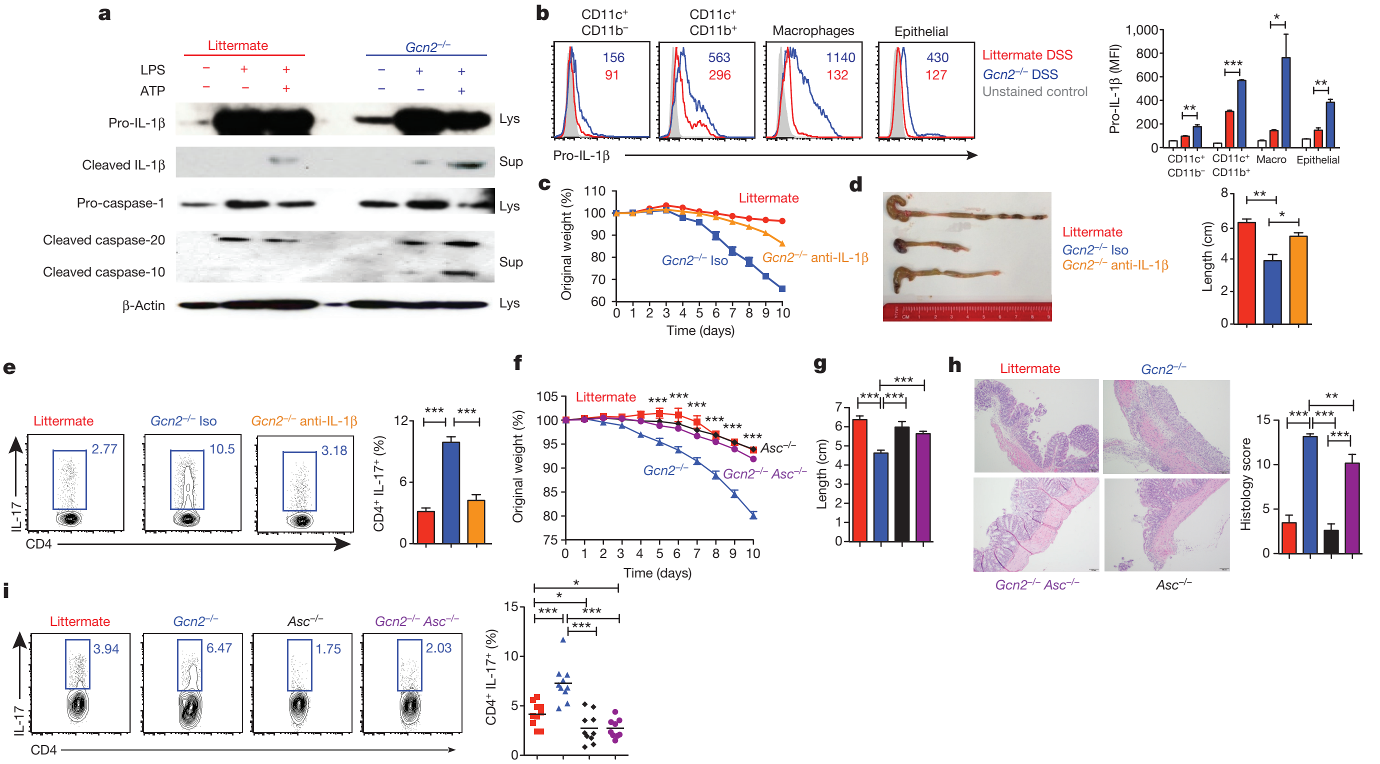


**Figure 3 | Enhanced intestinal inflammation in *Gcn2*$^{-/-}$ mice is dependent on inflammasome activation. a**, Western blot analysis of pro-IL-1β, pro-caspase, cleaved IL-1β and cleaved caspase in lysate and culture supernatants of bone marrow dendritic cell cultures from wild-type and *Gcn2*$^{-/-}$ mice treated with lipopolysaccharide (LPS) alone or LPS plus ATP (potassium efflux agent that triggers inflammasomes) under amino acid starvation conditions. **b**, Quantification of MFIs of pro-IL-1β in colonic APC subsets in *Gcn2*$^{-/-}$ and littermates after DSS. Macro, macrophages. **c**–**e**, Body weight (**c**), colon length (**d**), and colonic $T_H17$ responses (**e**) in DSS-treated *Gcn2*$^{-/-}$ mice treated with anti-IL-1β antibody or isotype control (Iso). **f**–**i**, Comparison of body weight (**f**), colon length (**g**), histology (**h**) and $T_H17$ frequencies (**i**) between littermates, *Gcn2*$^{-/-}$, *Asc*$^{-/-}$ (*Asc* is also known as *Pycard*) and *Gcn2*$^{-/-}$*Asc*$^{-/-}$ mice subjected to DSS-induced colitis. Data are from one experiment that is representative of three separate experiments ($n = 4$–$5$). $^*P < 0.05$; $^{**}P < 0.005$, $^{***}P < 0.0005$. Error bars indicate mean ± s.e.m.

protein-restricted diets and normal diets had similar weights before DSS (data not shown). After DSS, wild-type mice on protein-restricted diets weighed significantly less than those on control diet (Fig. 4c), but this was not the case in *Gcn2*$^{-/-}$ mice, indicating that GCN2 protected against gut inflammation (Fig. 4c). The colon lengths and histopathology (epithelial integrity, cellular infiltration, crypt loss) were similar (Extended Data Fig. 8e and data not shown). By contrast, mice on protein-modified diets showed a reduced incidence of 'bloody diarrhoea', compared with control diet mice (Fig. 4d). Remarkably, the frequencies of colonic $T_H17$ cells were significantly lower in wild-type mice on modified diets compared with mice on control diets (Fig. 4e). In contrast, there were no significant differences in the $T_H1$ responses or regulatory T cells, indicating that amino-acid-restricted diets selectively impair $T_H17$ responses (Extended Data Fig. 8f). Notably, there were no differences in the $T_H17$ frequencies among *Gcn2*$^{-/-}$ mice on various diets, indicating that this effect is GCN2 dependent (Fig. 4e). Together, these data demonstrate that amino acid starvation protects the symptoms of colitis and limits $T_H17$ cells via a GCN2-dependent mechanism.

We demonstrate that GCN2 suppresses intestinal inflammation and $T_H17$ responses via a mechanism dependent on autophagy and sequestration of ROS, which is a trigger for inflammasome activation (Extended Data Fig. 9a). *Gcn2*$^{-/-}$ mice displayed enhanced ROS and inflammasome activation, leading to increased inflammation and $T_H17$ responses (Figs 1–3). Thus, blockade of ROS and IL-1β led to lower inflammation and $T_H17$ responses in *Gcn2*$^{-/-}$ mice (Figs 2 and 3). In addition, *Gcn2*$^{-/-}$ mice were deficient in autophagy, which sequesters ROS, and consistent with this there was enhanced ROS and $T_H17$ inflammation in Atg5$^{\Delta APC}$ and Atg7$^{\Delta APC}$ mice (Fig. 1 and Extended Data Fig. 7). Future studies aimed at the functional reconstitution of a constitutively active autophagy pathway specifically in intestinal APCs and epithelial cells in *Gcn2*$^{-/-}$ mice should provide greater insight into the extent to which the observed phenotype in *Gcn2*$^{-/-}$ mice is due to impaired autophagy. Consistent with these results it is known that halofuginone, a compound that activates the amino acid starvation response, selectively inhibits mouse and human $T_H17$ differentiation[24]. Remarkably, we observed that a low protein diet, which activates the amino acid starvation response pathway, reduces the symptoms of colitis and colonic $T_H17$ responses. Although prolonged protein deficiency impairs critical immune functions[25], short-term protein restriction can enhance immunity to pathogens[26–28] and cancer[29]. Also, pharmacological activation of GCN2 protected mice against ischaemia reperfusion injury[30].

It is tempting to speculate on the evolutionary significance of coupling amino acid starvation with control of inflammation. Tissue injury and cell death, which occur during inflammation, inevitably result in tissue regeneration. Tissue regeneration, in turn, is accompanied by protein synthesis, which could lead to amino acid depletion in the cytosol. The consequent activation of GCN2 will suppress inflammasome activation through the mechanisms described here, in effect representing a negative feedback mechanism that limits the inflammation (Extended Data Fig. 9b). Our results show a role for GCN2 in protecting mice against intestinal inflammation. Thus, targeting the GCN2 pathway may provide new strategies for pharmacological intervention for the amelioration of inflammatory bowel disease and other inflammatory disorders.

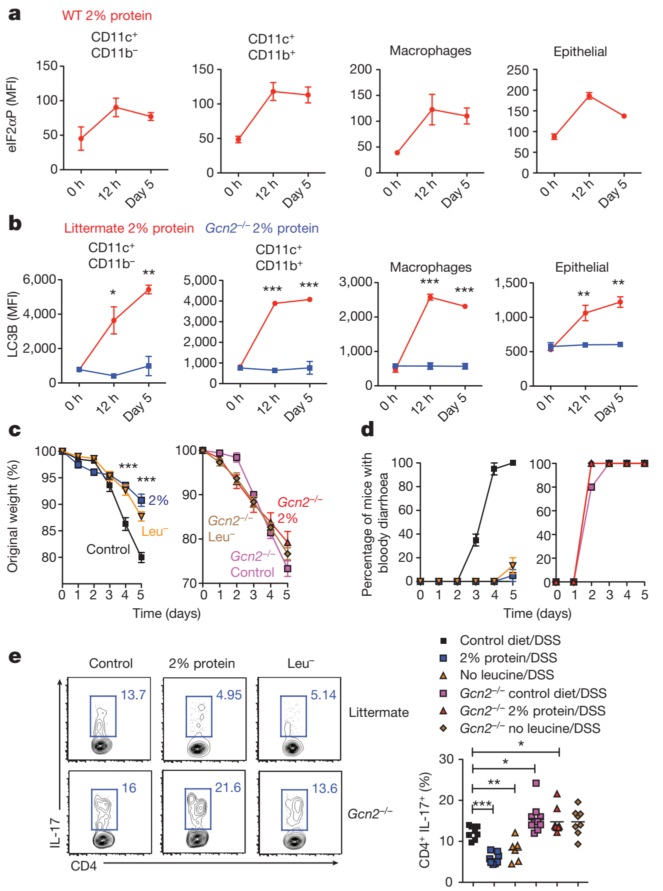


**Figure 4 | Dietary restriction of amino acids can partially protect against DSS-induced colitis. a**, Expression of p-eIF2α on APC subsets and epithelial cells from wild-type (WT) mice on a 2% protein diet. **b**, Kinetics of LC3B–GFP expression after a 2% protein diet in colonic APC subsets and epithelial cells isolated from *Gcn2*$^{-/-}$ LC3–GFP and LC3–GFP mice. **c**, Mice on a modified protein diet are protected from DSS-induced colitis. **c**–**e**, Weight loss (**c**), percentage of animals with bloody diarrhoea (**d**) and colonic $T_H17$ responses (**e**) in 3% DSS-induced wild-type or *Gcn2*$^{-/-}$ mice that were on protein-modified diets (2% protein diet or leucine-deficient diet) compared with control diet (16%). Data are from two separate experiments that were then pooled. $*P < 0.05$; $**P < 0.005$, $***P < 0.0005$. Error bars indicate mean ± s.e.m.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 5 November 2015; accepted 29 January 2016.**

**Published online 16 March 2016.**

1. Zhou, R., Yazdi, A. S., Menu, P. & Tschopp, J. A role for mitochondria in NLRP3 inflammasome activation. *Nature* **469,** 221–225 (2011).
2. Kawai, T. & Akira, S. Toll-like receptors and their crosstalk with other innate receptors in infection and immunity. *Immunity* **34,** 637–650 (2011).
3. Pulendran, B. The varieties of immunological experience: of pathogens, stress, and dendritic cells. *Annu. Rev. Immunol.* **33,** 563–606 (2015).
4. Donnelly, N., Gorman, A. M., Gupta, S. & Samali, A. The eIF2α kinases: their structures and functions. *Cell. Mol. Life Sci.* **70,** 3493–3511 (2013).
5. Han, A. P. *et al.* Heme-regulated eIF2α kinase (HRI) is required for translational regulation and survival of erythroid precursors in iron deficiency. *EMBO J.* **20,** 6909–6918 (2001).
6. Querec, T. D. *et al.* Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nature Immunol.* **10,** 116–125 (2009).
7. Ravindran, R. *et al.* Vaccine activation of the nutrient sensor GCN2 in dendritic cells enhances antigen presentation. *Science* **343,** 313–317 (2014).
8. Funke, B. *et al.* Functional characterisation of decoy receptor 3 in Crohn's disease. *Gut* **58,** 483–491 (2009).
9. Kugathasan, S. *et al.* Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nature Genet.* **40,** 1211–1215 (2008).
10. Kleene, S. J., Toews, M. L. & Adler, J. Isolation of glutamic acid methyl ester from an *Escherichia coli* membrane protein involved in chemotaxis. *J. Biol. Chem.* **252,** 3214–3218 (1977).
11. Harding, H. P., Zhang, Y. & Ron, D. Protein translation and folding are coupled by an endoplasmic-reticulum-resident kinase. *Nature* **397,** 271–274 (1999).
12. Back, S. H. *et al.* Translation attenuation through eIF2α phosphorylation prevents oxidative stress and maintains the differentiated state in β cells. *Cell Metab.* **10,** 13–26 (2009).
13. Cao, S. S. *et al.* Phosphorylation of eIF2α is dispensable for differentiation but required at a posttranscriptional level for paneth cell function and intestinal homeostasis in mice. *Inflamm. Bowel Dis.* **20,** 712–722 (2014).
14. Tattoli, I. *et al.* Amino acid starvation induced by invasive bacterial pathogens triggers an innate host defense program. *Cell Host Microbe* **11,** 563–575 (2012).
15. Saitoh, T. *et al.* Loss of the autophagy protein Atg16L1 enhances endotoxin-induced IL-1β production. *Nature* **456,** 264–268 (2008).
16. Mizushima, N. & Kuma, A. Autophagosomes in GFP-LC3 transgenic mice. *Methods Mol. Biol.* **445,** 119–124 (2008).
17. Martinez, J. *et al.* Molecular characterization of LC3-associated phagocytosis reveals distinct roles for Rubicon, NOX2 and autophagy proteins. *Nature Cell Biol.* **17,** 893–906 (2015).
18. Damiani, C. R. *et al.* Oxidative stress and metabolism in animal model of colitis induced by dextran sulfate sodium. *J. Gastroenterol. Hepatol.* **22,** 1846–1851 (2007).
19. Brubacher, J. L. & Bols, N. C. Chemically de-acetylated 2′,7′-dichlorodihydrofluorescein diacetate as a probe of respiratory burst activity in mononuclear phagocytes. *J. Immunol. Methods* **251,** 81–91 (2001).
20. Julian, D., April, K. L., Patel, S., Stein, J. R. & Wohlgemuth, S. E. Mitochondrial depolarization following hydrogen sulfide exposure in erythrocytes from a sulfide-tolerant marine invertebrate. *J. Exp. Biol.* **208,** 4109–4122 (2005).
21. Zhang, P. *et al.* The GCN2 eIF2α kinase is required for adaptation to amino acid deprivation in mice. *Mol. Cell. Biol.* **22,** 6681–6688 (2002).
22. Hao, S. *et al.* Uncharged tRNA and sensing of amino acid deficiency in mammalian piriform cortex. *Science* **307,** 1776–1778 (2005).
23. Anthony, T. G. *et al.* Preservation of liver protein synthesis during dietary leucine deprivation occurs at the expense of skeletal muscle mass in mice deleted for eIF2 kinase GCN2. *J. Biol. Chem.* **279,** 36553–36561 (2004).
24. Sundrud, M. S. *et al.* Halofuginone inhibits $T_H17$ cell differentiation by activating the amino acid starvation response. *Science* **324,** 1334–1338 (2009).
25. Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474,** 327–336 (2011).
26. Hu, J. F. *et al.* Repression of hepatitis B virus (HBV) transgene and HBV-induced liver injury by low protein diet. *Oncogene* **15,** 2795–2801 (1997).
27. Ariyasinghe, A. *et al.* Protection against malaria due to innate immunity enhanced by low-protein diet. *J. Parasitol.* **92,** 531–538 (2006).
28. Oarada, M. *et al.* Beneficial effects of a low-protein diet on host resistance to *Paracoccidioides brasiliensis* in mice. *Nutrition* **25,** 954–963 (2009).
29. Li, C. *et al.* Immunopotentiation of NKT cells by low-protein diet and the suppressive effect on tumor metastasis. *Cell. Immunol.* **231,** 96–102 (2004).
30. Peng, W. *et al.* Surgical stress resistance induced by single amino acid deprivation requires *Gcn2* in mice. *Sci. Transl. Med.* **4,** 118ra11 (2012).

**Acknowledgements** This work was supported by grants from the US National Institutes of Health (NIH; grants R37 DK057665, R37 AI048638, U19 AI090023, U19 AI057266) and from the Bill & Melinda Gates Foundation to B.P. R.J.K. was supported by NIH grants R01 DK088227, R37 DK042394, R01 DK103185 and the Crohn's & Colitis Foundation of America Senior Fellow Award 3800. We thank B. Cervasi and K. P. Gill for help with sorting and D. Levesque, D. Hampton, S. Weismann and colleagues for animal husbandry and veterinary support. We thank the Center for AIDS Research core at Yerkes Primate Research Center for help with qPCR (grant P30-AI-50409) and the Integrated Cellular Imaging Center at Winship Cancer Center, Emory University for help with confocal microscopy. We also thank the Yerkes Pathology Core for performing the immunofluorescence and histology analysis. We thank S. Virgin for his advice and review of the autophagy experiments.

**Author Contributions** R.R., J.L. and B.P. designed experiments and wrote the manuscript. R.R., J.L., N.K., D.K.M., H.M., S.L. and B.L. conducted the experiments. H.I.N. and L.G. performed bioinformatics analysis of the public databases. P.H. and Y.-c.W. genotyped mice. P.S. performed the histology analysis. J.M. provided critical insight and advice about the autophagy experiments. R.J.K. provided reagents and edited the manuscript.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to B.P. (bpulend@emory.edu).

## METHODS

**Mice.** *Gcn2*$^{-/-}$ mice[31] were provided by D. Munn. The *Gcn2*-floxed mice[32] were obtained from Jackson Laboratories, and bred onsite. *Perk*-floxed mice were obtained from D. Cavener. To delete specifically GCN2 or PERK in APCs, floxed *Gcn2* mice (*Gcn2*$^{fl/fl}$) were crossed with *Cd11c cre* or villin *cre* mice that express the Cre enzyme under the control of the CD11c promoter[33] or villin promotor[34], respectively, generating *Gcn2*$^{\Delta\mathrm{APC}}$, *Gcn2*$^{\Delta\mathrm{villin}}$, *Perk*$^{\Delta\mathrm{APC}}$ and *Perk*$^{\Delta\mathrm{villin}}$ mice. eIF2α-Ser51Ala Tg (*Eif2a* floxed) mice were provided by R.J.K.[12], and crossed to CD11c-*cre* and villin-*cre* mice to obtain the conditional expression of non-phosphorylatable Ser51Ala mutant eIF2α in APCs or intestinal epithelial cells. Successful Cre-mediated deletion was confirmed by PCR and protein expression (fluorescence-activated cell sorting (FACS) or western blot) (Extended Data Fig. 10). LC3–GFP mice were generated by N. Mizushima[35] and provided by H. Virgin. These mice were crossed with *Gcn2*$^{-/-}$ mice to generate *Gcn2*$^{-/-}$ LC3–GFP mice.

Animal studies were conducted using age-matched littermate controls for each experiment. Both male and female mice were used and were between 8 and 14 weeks of age at the time of experiments. Mice were maintained under specific-pathogen-free conditions in the Emory Vaccine Center vivarium. All animal protocols were reviewed and approved by the Institutional Animal Care and Use Committee of Emory University. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments.

**Histology.** Formalin-fixed and paraffin-embedded murine intestinal tissue was sectioned (4 μm) and standard haematoxylin and eosin (H&E) stains were carried out to assess DSS-induced damage. Paraffin-embedded sections were subjected to deparaffinization in xylene, rehydration in graded series of ethanol, and rinsing with distilled water. The resident pathologist at Yerkes performed all histopathology analyses in a blinded manner.

Immunohistochemistry was performed by using labelled anti-mouse/anti-rabbit/anti-goat biotin antibody followed by streptavidin–alkaline phosphatase using a multistep staining protocol. All reactions were detected by development of the chromogen (Warp Red; Biocare Medical). Appropriate positive and negative controls were run in parallel. After rehydration, slides were treated with a target retrieval solution (DIVA decloaker; Biocare Medical) and then steamed for antigen retrieval for 20 min and cooled for 20 min. The following primary antibodies were used for immunohistochemical staining: Ki67 (#12202, Cell Signaling Technology, 1:200), lysozyme (Sc-27956, Santa Cruz, 1:50), chromagranin A (ab15160, Abcam, 1:200). Nuclei were counterstained using Gill's haematoxylin. Human sections from naive or inflamed colons (Pantomics, catalogue numbers COLO1 and COLO2) were stained in a similar fashion with the following primary antibodies: PERK (#T980, Cell Signaling Technology, 1:200), phospho-GCN2 (ab32026, Abcam, 1:200), PKR (ab32036, Abcam, 1:200), anti-IL-1β (ab9722, Abcam, 1:200).

**Isolation of intestinal APCs and lymphocytes.** For epithelial and lamina propria (LP) lymphocyte isolation, mice were euthanized and the intestine was washed, cleaned of fat tissue and Peyer's patches, and longitudinally cut and suspended in 1× HBSS with 20 mM HEPES, 1 mM dithiothreitol (DTT) and 5 mM EDTA for 30 min at 37 °C. After washing with 1× HBSS, pieces were digested with collagenase VIII (Sigma) (1 U ml$^{-1}$ in RPMI with 2% FCS) for 30 min at 37 °C with shaking (150 r.p.m.). Tissue was processed through a 100-μm cell strainer, and the resulting suspension was pelleted. For lymphocyte isolation, cells derived following collagenase were resuspended in 7 ml of 40% Percoll and layered on top of 2 ml of 70% Percoll (GE Amersham). After centrifugation for 15 min at 1,500 r.p.m. without brakes, the middle layer was removed, washed in 2% FBS in RPMI and the lymphocytes were obtained. For LP dendritic cells and macrophages, the collagenase-digested cells were filtered through a 100-μm strainer and pelleted and stained. For cell sorting, APCs in this preparation were enriched with CD11c$^{+}$ and CD11b$^{+}$ magnetic beads according to the manufacturer's instructions (Miltenyi Biotec) and sorted on FACS Aria at the Emory Vaccine Center flow cytometer core (BD Biosciences). For flow cytometry, cells were stained with a combination of the following fluorescence-conjugated monoclonal antibodies: phycoerythrin (PE)–Texas-red-conjugated anti-CD11b (Invitrogen), PE-Cy7-conjugated anti-F4/80 (eBioscience), allophycocyanin (APC)-conjugated anti-CD3 (eBioscience), Pac-Blue-conjugated anti-CD11c (eBioscience), eFluor650NC-conjugated anti-CD45 (eBioscience), PerCP-Cy5.5-conjugated anti-CD4 (eBioscience). For studying the levels of pro-IL-1β, cells were stained intracellularly with antibody from Affymetrix eBioscience (12-7114). Dead cells were excluded from analysis through the use of a LIVE/DEAD Fixable Aqua Dead Cell Stain kit (Invitrogen). Samples were acquired on an LSR II (BD Biosciences) and were analysed with FlowJo software (v.9.2; TreeStar).

***In vivo* intestinal barrier function assays.** Naive or 3-day DSS-treated wild-type or *Gcn2*$^{-/-}$ mice were fasted overnight, and FITC-dextran (0.6 mg g$^{-1}$; Sigma) diluted in PBS was gavaged the following day. Fluorescence intensity of plasma samples was measured (excitation 485 nm/emission 535 nm) 4 h after gavage.

***In vitro* lymphocyte co-culture.** FACS-sorted LP APC subsets ($1 \times 10^{5}$) were co-cultured with naive CD4$^{+}$CD62L$^{+}$ OT-II CD4$^{+}$ transgenic T cells ($1 \times 10^{5}$) and OVA peptide (ISQVHAAHAEINEAGR; 2 μg ml$^{-1}$) in a total volume of 200 μl RPMI complete medium. The culture supernatants were analysed after 72 h and cells were harvested and restimulated for 6 h with plate-bound antibodies against CD3 (10 μg ml$^{-1}$; 145.2C11 from Becton Dickinson) and CD28 (2 μg ml$^{-1}$; 37.51 from Becton Dickinson), in the presence of brefeldin A (Becton Dickinson) for intracellular cytokine detection (IL-17A and IFN-γ). For analysis of IL-17A and IFN-γ responses from freshly isolated LP, the Percoll-purified lymphocytes from colon or small intestine were stimulated with PMA (100 ng ml$^{-1}$) and ionomycin (1,000 ng ml$^{-1}$) for 4 h in the presence of GolgiPlug and then stained for intracellular IL-17A and IFN-γ for FACS analysis.

**Phospho-eIF2α staining by flow cytometry.** Mice were orally gavaged with 2% DSS in 200 μl volume. The LP cells were isolated and immediately fixed using the BD fixation solution for 10 min. Fixed cells were stained for intracellular p-eIF2α using a monoclonal antibody from Cell Signaling (3398). Cells were then stained with anti-rabbit IgG-PE (Jackson Laboratories, catalogue number 111-116-144) at a 1:400 dilution.

**Gene expression analysis.** Total RNA was extracted from gut samples using Qiagen RNeasy mini kit (catalogue number 74104) according to the manufacturer's recommended protocol with the column DNase–RNase-free treatment. Extracted RNA was reverse transcribed with SuperScript Vilo cDNA synthesis kit (catalogue 11754050) according to the manufacturer's protocol. An aliquot (10 ng) of cDNA was used to quantitate *Ang4* ((Mm03647554_g1), *Lyz* (Mm00727183_s1) and *Reg3g* (Mm00441127_m1) on the ABI7900 system. *GADPH* (Mm99999915_g1) was used as the reference gene. The comparative gene expression method was used to determine the relative quantitation.

**ROS and MitoSOX staining by flow cytometry.** For detecting the ROS levels, the LP preparations were incubated with CM-H2DCFDA (10 mM; C6827, Life Technologies), a cell-permeable indicator for ROS. After incubation the levels of fluorescence were measured by flow cytometry. For MitoSOX staining the LP preparation was incubated with MitoSOX (M36008, Life Technologies) for 15 min (manufacturer's instructions). Cells were then surface stained and analysed by flow cytometry.

**NAC and anti-IL-1β treatment *in vivo*.** For *in vivo* treatment, NAC (Sigma) was injected into mice intraperitoneally (i.p.; 275 mg kg$^{-1}$) in PBS solution, pH 7.4, every other day at days 1, 3, 5 and 7 after DSS. For *in vivo* IL-1β neutralizing experiments, 300 μg per mouse of anti-mouse IL-1β (BioXcell, catalogue number BE0246) was injected i.p. on days 1, 3, 5 and 7 after DSS. Control mice received 300 μg isotype control antibody (Hamster IgG, BioXcell BE0091) i.p.

**Detection of LC3–GFP.** Mice were euthanized, and intestines were fixed (3.7% formaldehyde for 3 min) and promptly washed with PBS. The tissue was fixed in formalin for an additional 12–18 h. Fixed tissue was embedded in OCT and sectioned on a cryotome into 6-μm sections. Slides were washed with PBS and mounted with Prolong Gold Antifade reagent with 4′,6-diamidino-2-phenylindole (DAPI; Invitrogen). Images of the sections were collected using LSM510 META confocal microscopy (Carl Zeiss). The *z*-stack images were collected and the GFP signals were analysed through the sections. Each individual crypt was analysed in three dimensions to reveal the number of LC3–GFP-positive crypts using an Imaris 7 3D/4D image processing and analysis software (Bitplane). Multiple crypts (5–7) were chosen from 3 different animals per group and the average mean LC3–GFP counts were quantified per crypt using Imaris software 7 (Bitplane).

***In vivo* autophagy flux.** To study autophagy flux choloroquine was used as a mode of inhibiting degradation of the components of the autophagosomes. Accumulated versus steady-state form of the autophagosomes in wild-type and *Gcn2*$^{-/-}$ mice after oral administration of DSS (200 μl of 3% DSS) was compared. Chloroquine (10 mg kg$^{-1}$) or saline was administered i.p. 2 h after DSS gavage and mice were killed at 12 h and 24 h after DSS treatment. The colons were cleaned, collagenase-treated and the LP preparation was quickly stained for various APC surface markers (CD11c, CD45, Epcam, MHC-II, CD11b) and intracellular levels of p62 (1:150) (H00008878-M01PE) and LC3B (1:150) (NB100-220F) to study the levels of intracellular autophagosomes. Intracellular staining was performed using BD perm buffer (BD Bioscience, 554723) after fixation with the BD Cytofix/Cytoperm solution (BD Bioscience, 554714). In some instances the LP was treated with digitonin (200 μg ml$^{-1}$) for 10 min before undergoing intracellular staining.

**Cytokine ELISA detection.** IL-17 and IL- 1β in culture supernatants were quantitated by ELISA according to product protocol (R&D Systems catalogue number DY42; BD Biosciences catalogue number 559603).

**DSS-induced colitis.** Acute colitis was induced by adding 2% DSS (MW 36,000–50,000; MP Biomedicals) to drinking water *ad libitum* for 7 consecutive days.

After the treatment, mice were kept with normal drinking water for 3 more days. Mice were killed and tissues analysed for immune and histological analysis. In addition, mice were orally gavaged with 200 μl of 2% DSS, thus minimizing the variation within a group owing to differences in the consumption of the drinking water. In the experiment involving protein restriction diets, wild-type (littermate) or *Gcn2*$^{-/-}$ mice were fed with 3% DSS to induce colitis in the wild-type mice. Three per cent, rather than 2% DSS was used to induce intestinal inflammation in wild-type mice, to be able to detect any effect of a low protein diet in these mice.

**Low protein mouse diet.** The mice were fed with either test diet 5CC7 (1812281), Baker amino-acid-defined diet (16% protein) or diet 5CC7 modified with 2% defined protein or no leucine (Test Diet Land O'Lakes Purina Feed).

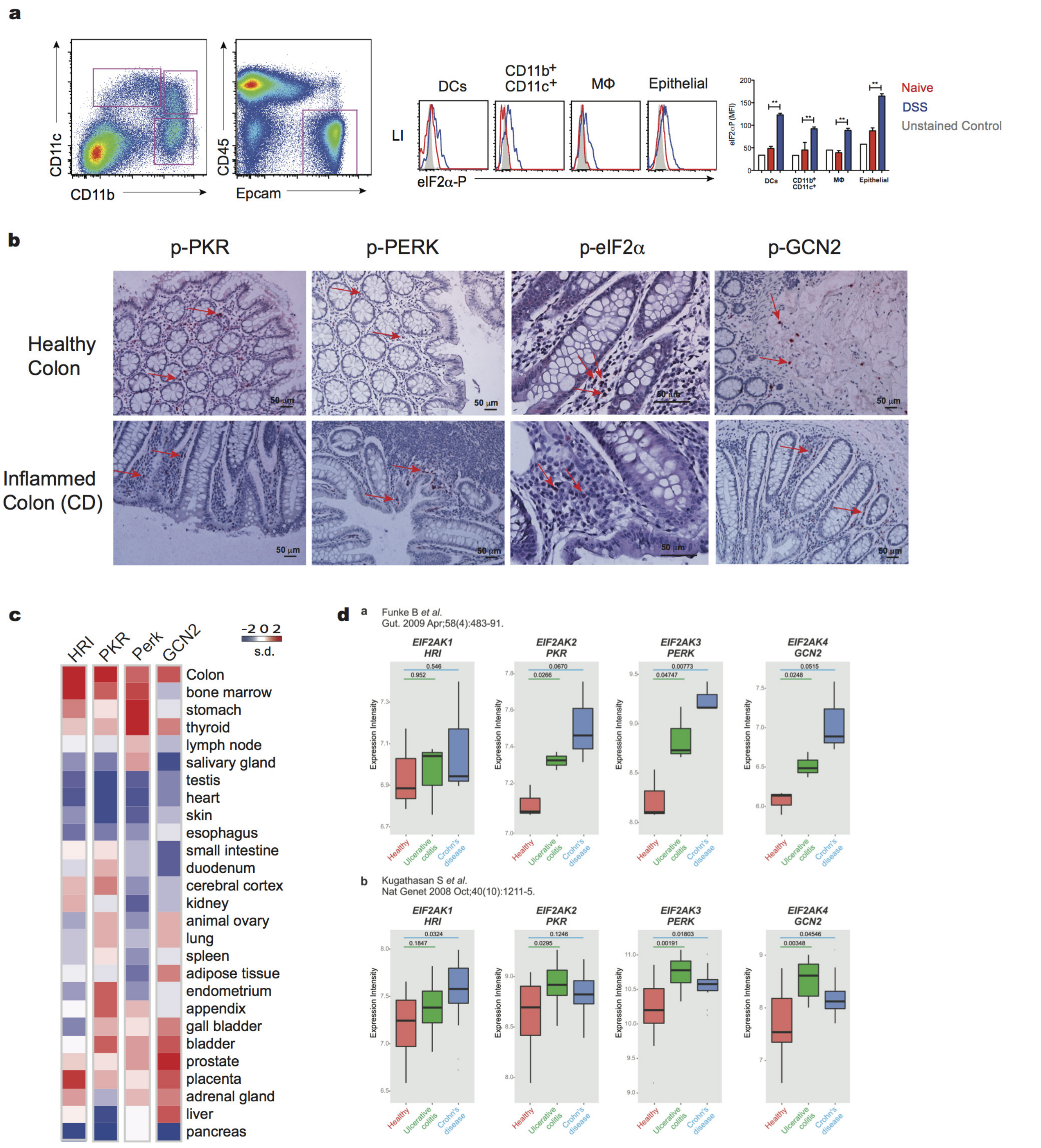
**Generation of murine bone-marrow-derived dendritic cells.** Mice tibiae and femurs were flushed with ice-cold PBS through a 70-μm cell strainer. Cells were pelleted and plated at a density of 5–7 × $10^6$ bone marrow cells per 10-cm Petri dish in (RPMI complete) in the presence of granulocyte–macrophage colony-stimulating factor (GM-CSF; 20 ng ml$^{-1}$, Peprotech) and IL-4 (5 ng ml$^{-1}$). At day 3, the cultures were supplemented with another 10 ml of RPMI complete plus GM-CSF. At day 6, bone-marrow-derived dendritic cells were harvested by gently flushing the cells from the plate.

***In vitro* activation of dendritic cells.** The bone-marrow-derived dendritic cells purified from wild-type or *Gcn2*$^{-/-}$ mouse tibia cultures were plated in 96-well microculture plates (1 × $10^5$ cells per well) and primed for 8 h with 100 ng ml$^{-1}$ LPS (InvivoGen, LPS-SM) in RPMI 1640 (US Biological; R8999-04A) plus 1% dialysed FBS (Gibco 26400-036). Cultures were further stimulated for 1 h with ATP (5 mM) (InvivoGen, tlrl-atp). After the stimulation, cell supernatants were collected and assayed for IL-1β by ELISA or western blot. Cell pellets were lysed and assayed for the presence of pro-IL-1β and pro-caspase 1.

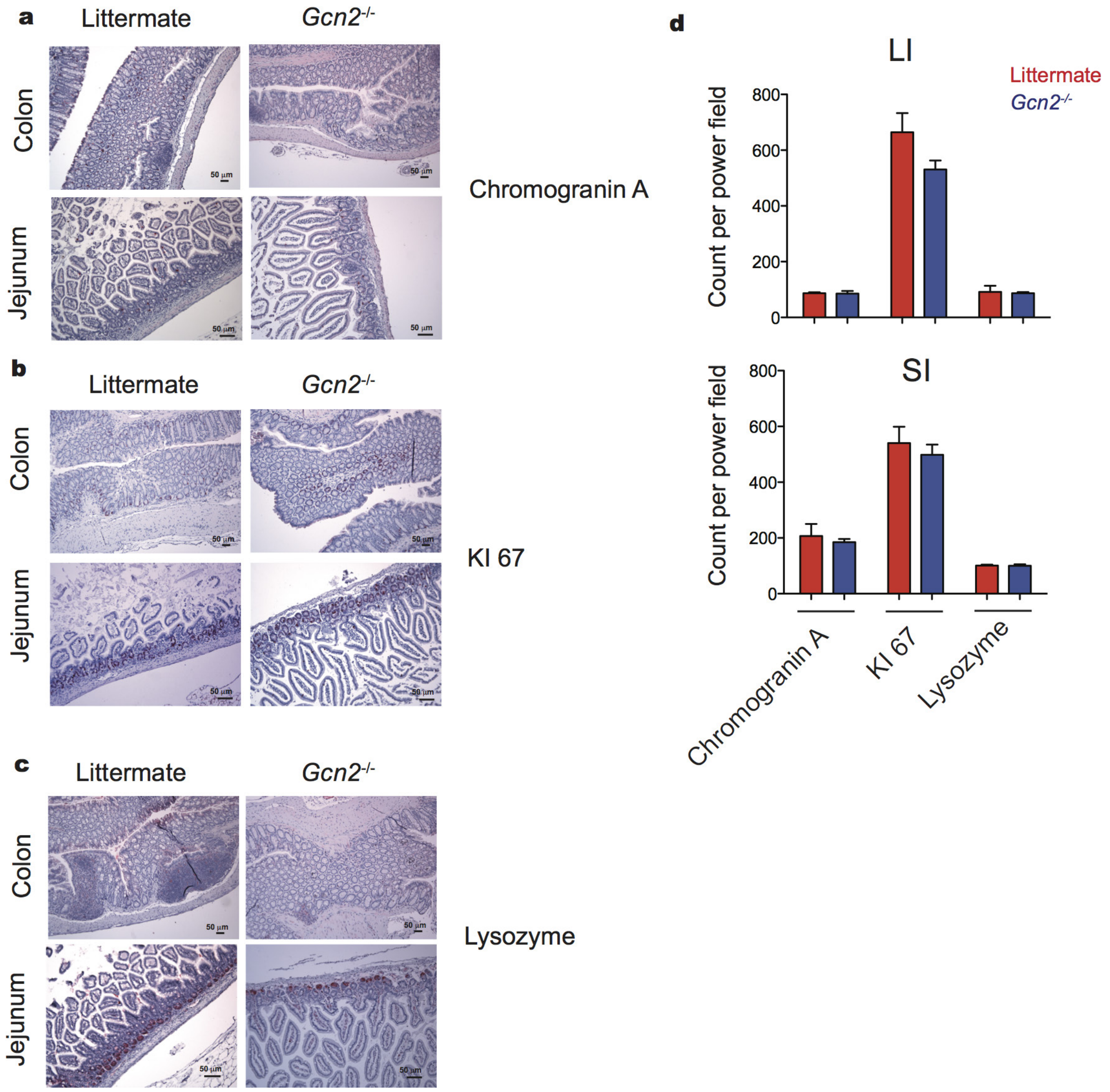
**Western blotting.** Purified dendritic cells (1 × $10^6$) were lysed with 100-μl protein extraction reagent (89900, Thermo Scientific) containing protease inhibitor (5872S, Cell Signaling). Equal amounts of protein (lysate) or supernatants were run on an SDS–PAGE and transferred onto nitrocellulose membranes after electro blotting. After blocking with 5% fat-free milk, the membranes were incubated at 4 °C overnight with the following primary antibodies: anti-mouse pro- and cleaved IL-1β (Cell Signaling, 12507), pro-caspase-1 (Abcam, ab108362), caspase-p20 and 10 (Adipogen, AG-20B-0042) as per the manufacturer's instructions. The membranes were then washed was incubated with horseradish-peroxidase-conjugated secondary antibody (Cell Signaling). Proteins were visualized with SuperSignal West Pico chemiluminescent substrate (34078, Thermo Scientific).

**Statistical analysis.** To assess the significance of a difference between groups, a two-sample, unpaired *t*-test was performed using Graph Prism software. A *P* value less than 0.05 was considered to be significant, a *P* value less than 0.01 was considered to be very significant, and a *P* value less than 0.001 was considered to be extremely significant.

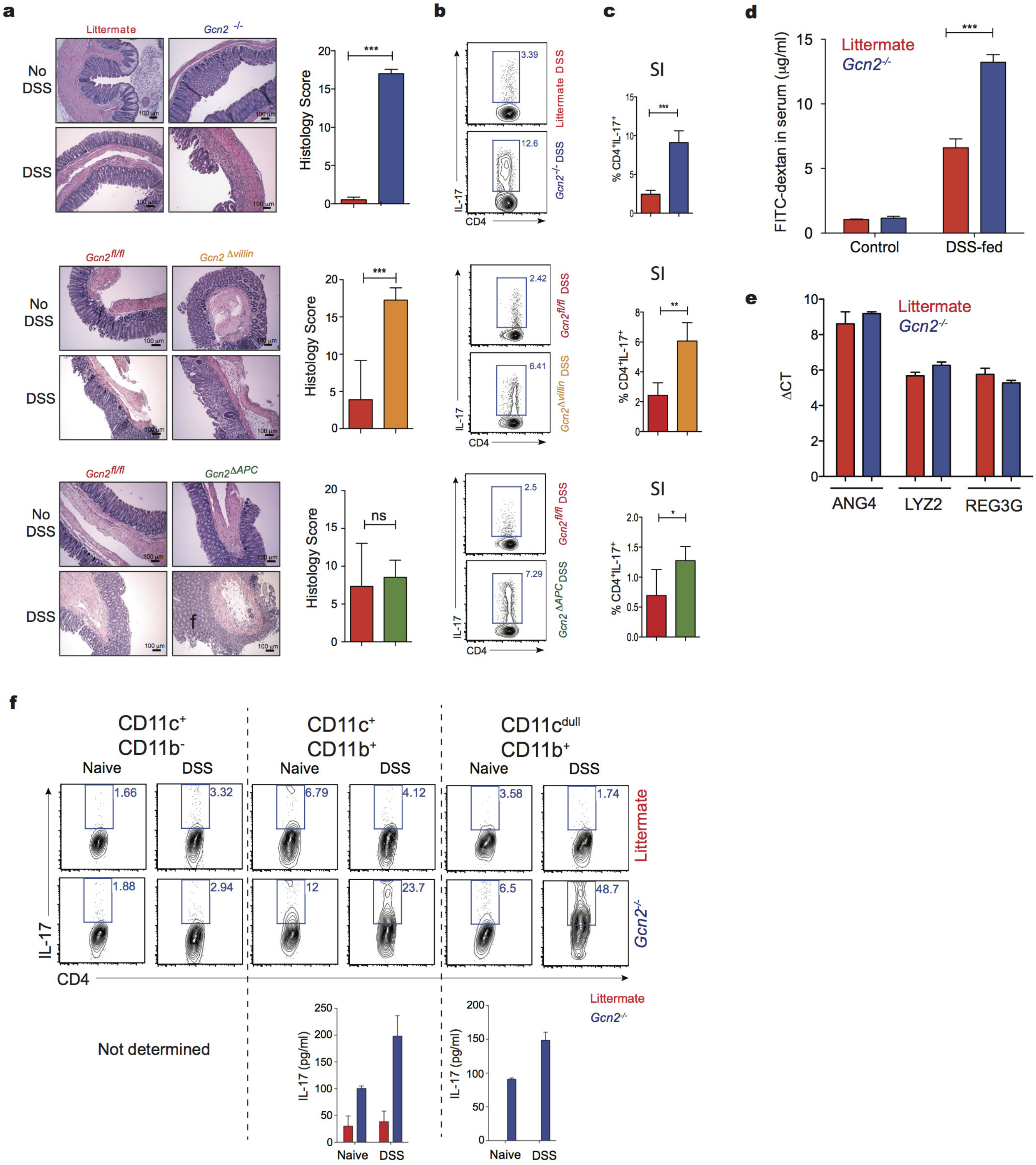
31. Harding, H. P. *et al.* Regulated translation initiation controls stress-induced gene expression in mammalian cells. *Mol. Cell* **6,** 1099–1108 (2000).
32. Zhang, P. *et al.* The PERK eukaryotic initiation factor 2 α kinase is required for the development of the skeletal system, postnatal growth, and the function and viability of the pancreas. *Mol. Cell. Biol.* **22,** 3864–3874 (2002).
33. Caton, M. L., Smith-Raska, M. R. & Reizis, B. Notch–RBP-J signaling controls the homeostasis of CD8$^{-}$ dendritic cells in the spleen. *J. Exp. Med.* **204,** 1653–1664 (2007).
34. Madison, B. B. *et al. Cis* elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J. Biol. Chem.* **277,** 33275–33283 (2002).
35. Mizushima, N., Yamamoto, A., Matsui, M., Yoshimori, T. & Ohsumi, Y. *In vivo* analysis of autophagy in response to nutrient starvation using transgenic mice expressing a fluorescent autophagosome marker. *Mol. Biol. Cell* **15,** 1101–1111 (2004).



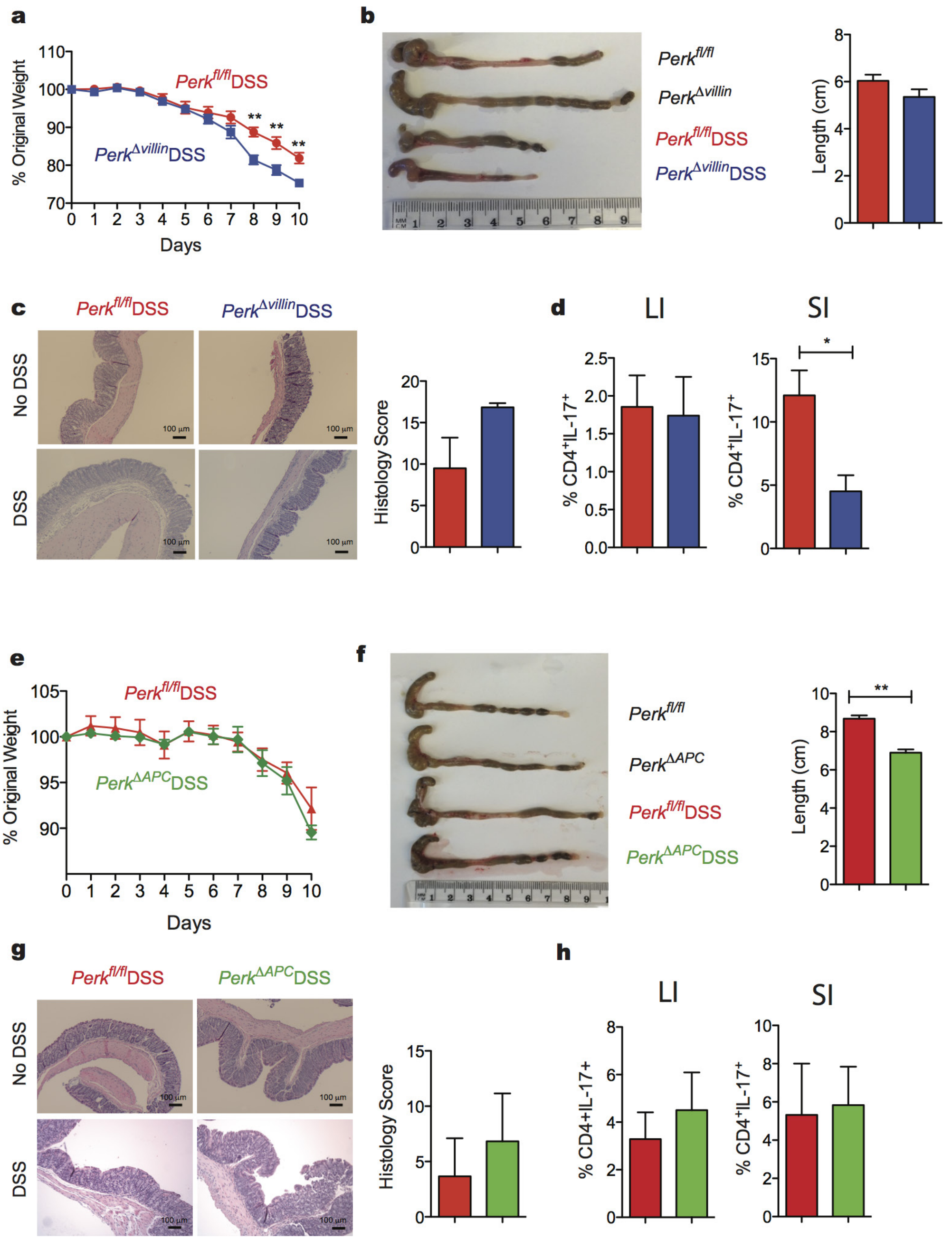
**Extended Data Figure 1 | eIF2α kinases are expressed in human and murine gut cells. a**, Analysis of p-eIF2α expression in APCs and epithelial cells in large intestine of naive and 2% DSS-treated mice by flow cytometry. DCs, dendritic cells; LI, large intestine. **b**, Comparison of immunohistological analysis of phosphorylated PKR, PERK, eIF2α and GNC2 in healthy and inflamed human colon tissue ($n = 1$). **c**, Expression levels of HRI, PKR, PERK and GCN2 in human organs quantified based on information from a public database (http://www.ebi.ac.uk). **d**, Expression intensity of various eIF2α kinases plotted from known published microarray data from colonic biopsies of patients with either ulcerative colitis or Crohn's disease compared to healthy controls. Data are from one experiment that is representative of three separate experiments. $*P < 0.05$, $**P < 0.005$, $***P < 0.0005$. Error bars indicate mean ± s.e.m.



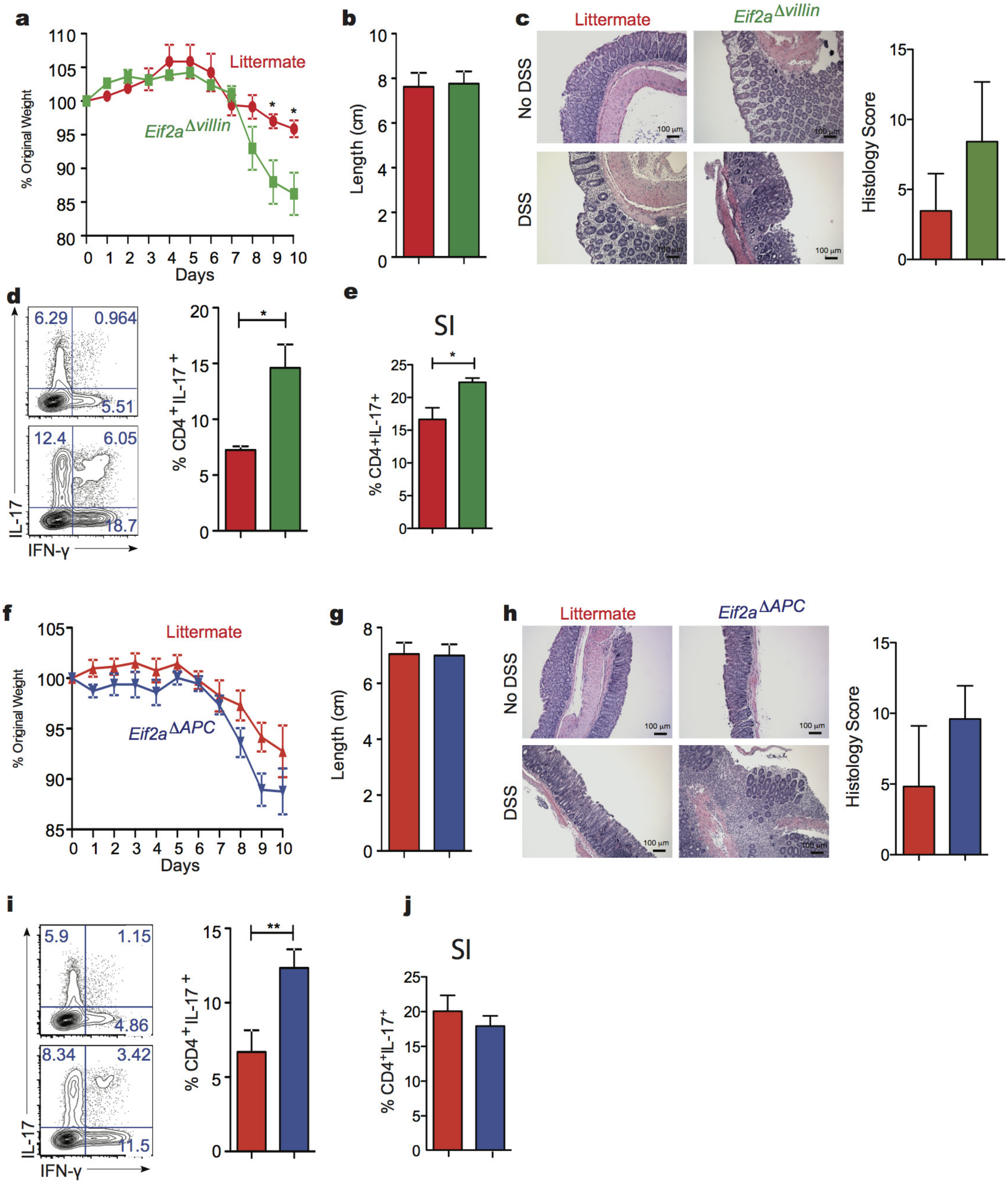
**Extended Data Figure 2 | GCN2 deficiency does not affect the proliferation or differentiation of intestinal epithelial cells.** **a**–**d**, Immunohistology analysis (**a**–**c**) and quantification (**d**) of colons and jejenums from *Gcn2*$^{-/-}$ and wild-type (littermate) mice for chromagranin A (**a**), Ki67 (**b**) and lysozyme (**c**). KO, knockout; WT, wild type. Data are from one experiment that is representative of three separate experiments ($n = 3$). $*P < 0.05$, $**P < 0.005$, $***P < 0.0005$. Error bars indicate mean ± s.e.m.



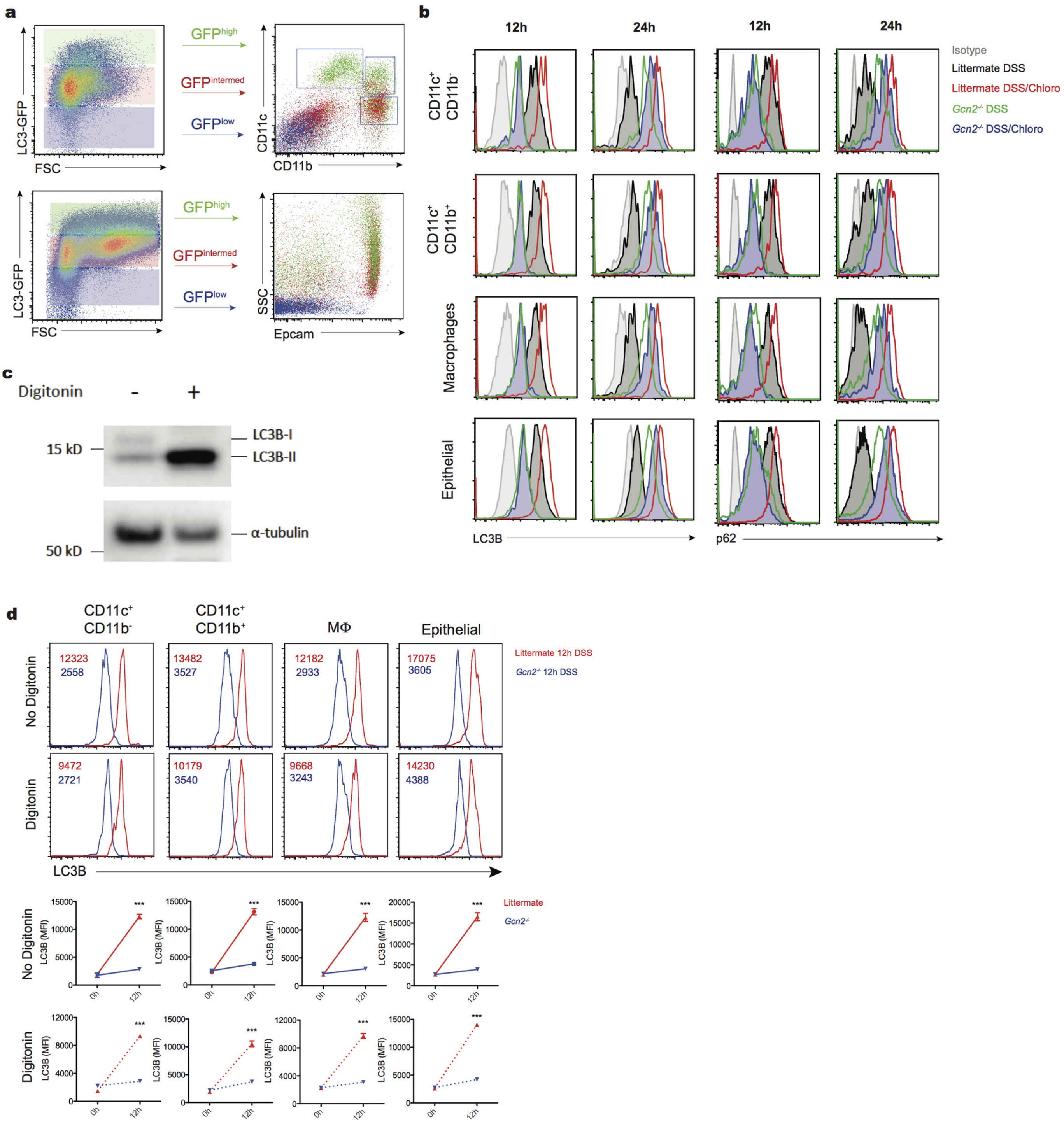
**Extended Data Figure 3 | GCN2 expression protects mice from DSS-induced colitis. a**, H&E staining of colon sections before and after DSS in wild-type versus $Gcn2^{-/-}$ mice, $Gcn2^{fl/fl}$ versus $Gcn2^{\Delta villin}$ and $Gcn2^{fl/fl}$ versus $Gcn2^{\Delta APC}$. **b**–**d**, IL-17 levels in large intestinal (**b**) and small intestinal (**c**) $CD4^+$ T cells measured by flow cytometry; $Gcn2^{-/-}$ mice show increased intestinal permeability after DSS treatment as evidenced by higher levels of fluorescein isothiocyanate (FITC)-conjugated dextran in the serum (**d**). SI, small intestine. **e**, Expression of antimicrobial defensins in wild-type and $Gcn2^{-/-}$ mice via quantitative polymerase chain reaction (qPCR). **f**, IL-17 production by flow cytometry and enzyme-linked immunosorbent assay (ELISA) of OTII-CD4 T cells after culturing with different large intestinal APC subsets. Data are from one experiment that is representative of three separate experiments ($n = 4$–$5$). $^*P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0005$. Error bars indicate mean ± s.e.m.



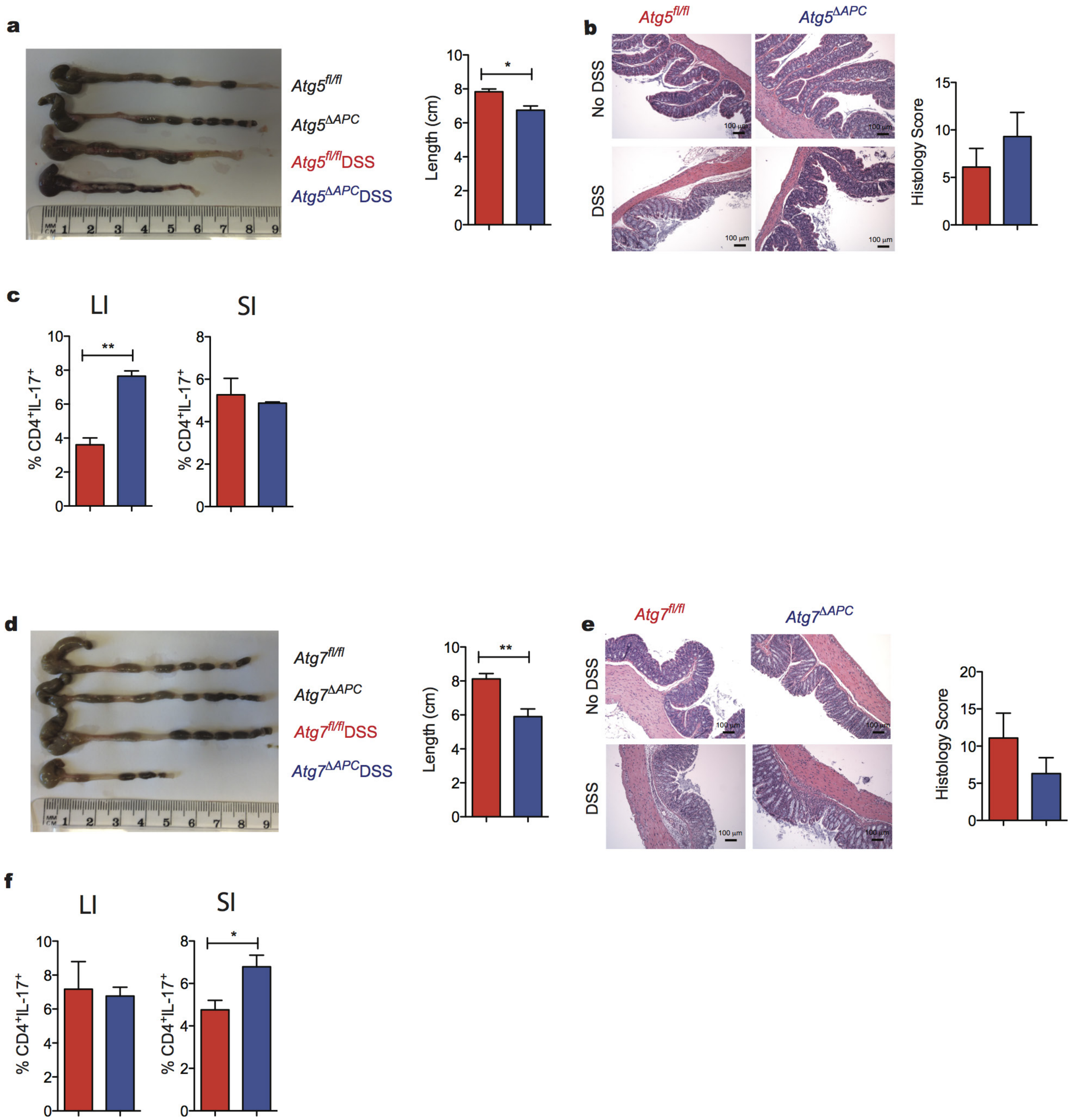
**Extended Data Figure 4 | PERK expression in epithelial cells has only a minor role in controlling mucosal homeostasis after DSS challenge.** **a**–**d**, The body weight (**a**), colon length (**b**), histology by H&E (**c**) and $T_H17$ responses (**d**) both in the colon (large intestine; LI) and small intestine (SI) of $Perk^{\Delta villin}$ and control wild-type littermates treated with DSS. **e**–**h**, The body weight (**e**), colon length (**f**), histology by H&E and histology score (**g**), and $T_H17$ responses (**h**) both in the colon (LI) and small intestine (SI) of $Perk^{\Delta APC}$ and control wild-type littermates treated with DSS. Data are representative of two separate experiments ($n = 5$). $*P < 0.05$; $**P < 0.005$, $***P < 0.0005$. Error bars indicate mean ± s.e.m.



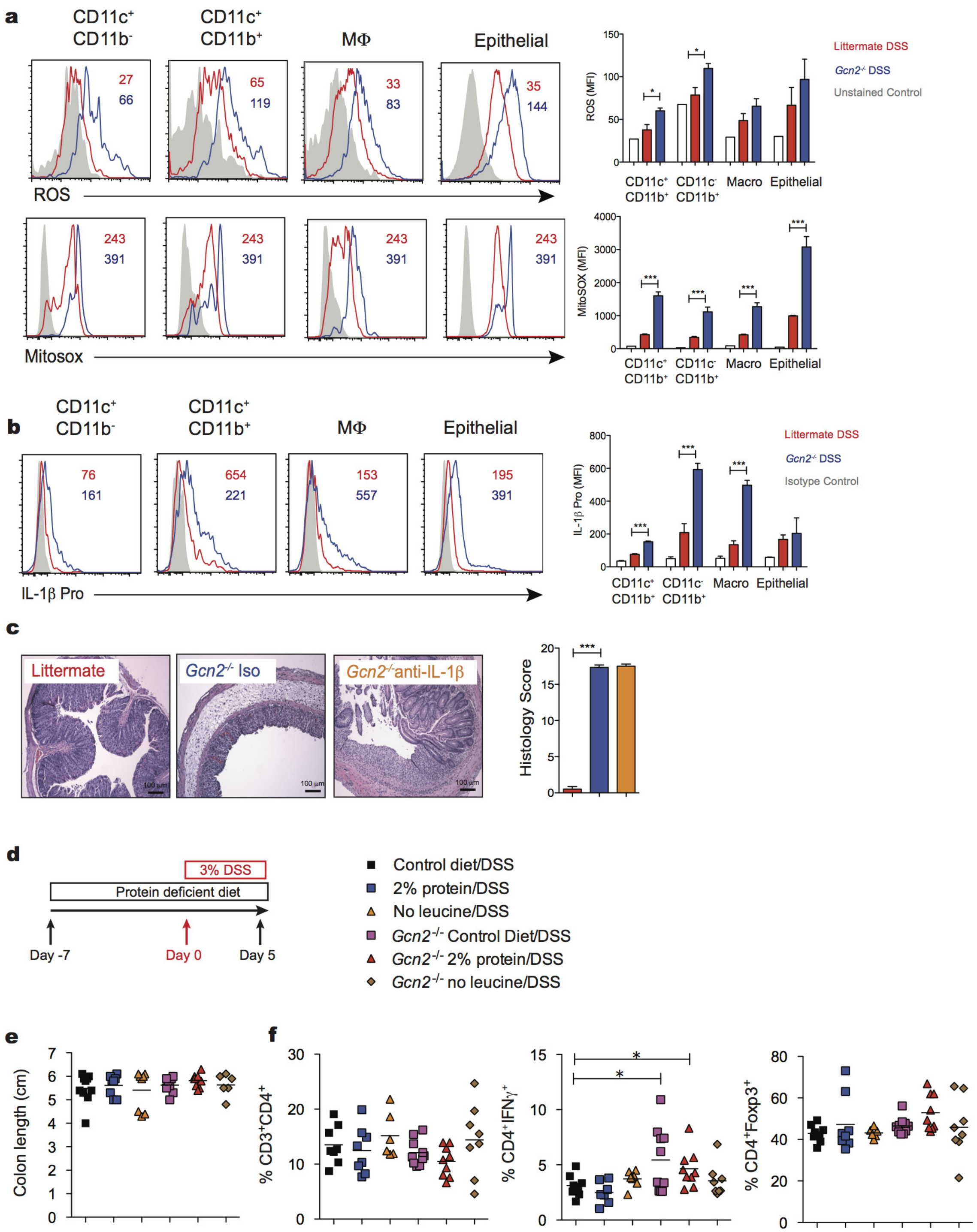
**Extended Data Figure 5 | eIF2α expression in epithelial cells and APCs control weight loss and partially control T$_H$17 responses after DSS challenge. a–e**, The body weight (**a**), colon length (**b**), histology by H&E and histology score (**c**), and T$_H$17 responses in both the colon (LI) (**d**) and small intestine (SI) (**e**) of *Eif2a*$^{\Delta\text{villin}}$ and control wild-type littermates treated with DSS. **f–j**, The body weight (**f**), colon length (**g**), histology by H&E and histology score (**h**), and T$_H$17 responses in both the colon (LI) (**i**) and small intestine (SI) (**j**) of *Eif2a*$^{\Delta\text{APC}}$ and control wild-type littermates treated with DSS. Data are representative of three separate experiments ($n = 5$). *$P < 0.05$, **$P < 0.005$, ***$P < 0.0005$. Error bars indicate mean ± s.e.m.



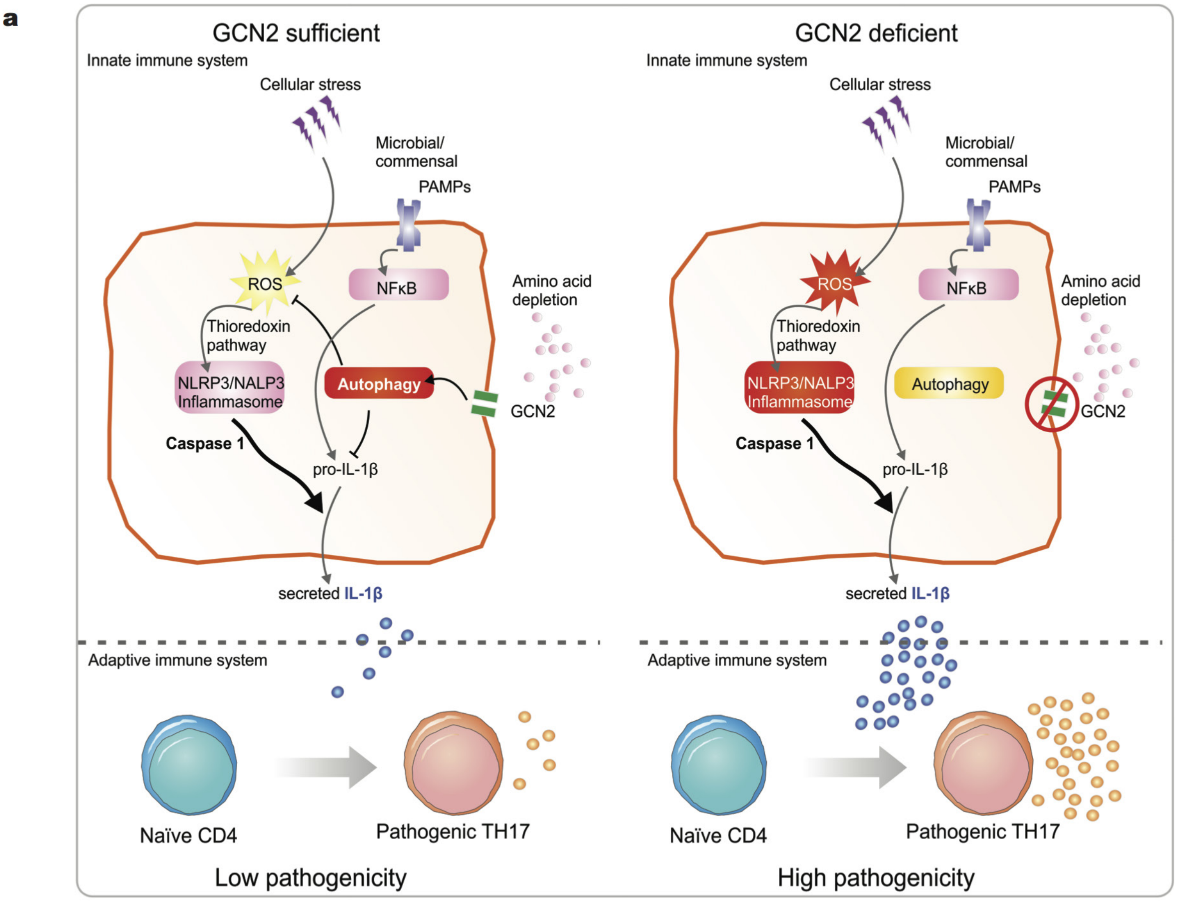
**Extended Data Figure 6 | Intestinal APCs and epithelial cells reveal high expression of LC3. a**, Expression of LC3–GFP in APC subsets and epithelial cells of naive LC3–GFP mice by flow cytometry ($n = 3$). Data are from a single experiment. **b**, Kinetic MFI comparison of LC3B and p62 expression with and without chloroquine on individual APC subsets and epithelial cells by flow cytometry after 12 and 24 h of single DSS administration. **c**, Western blot detection of LC3-I and II on lamina propria APCs before and after digitonin. **d**, LC3B staining of individual APCs and epithelial cells 12 h after they were treated with DSS. The portion of the lamina propria cells were subjected to digitonin before intracellular staining with the LC3B antibody. Data are representative of two separate experiments ($n = 4$–5). *$P < 0.05$, **$P < 0.005$, ***$P < 0.0005$. Error bars indicate mean ± s.e.m.

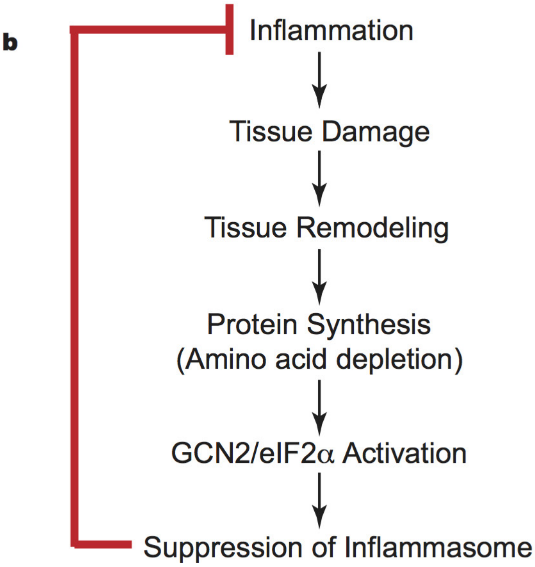


**Extended Data Figure 7 | Atg5 and Atg7 expression in APCs partially protects mice from DSS challenge. a–c**, Colon length (**a**), histology by H&E and histology score (**b**), and $T_H17$ responses (**c**) in both the colon (LI) and small intestine (SI) of $Atg5^{\Delta APC}$ and control wild-type littermates treated with DSS. **d–f**, Colon length (**d**), histology by H&E and histology score (**e**), and $T_H17$ responses (**f**) in both the colon (LI) and small intestine (SI) of $Atg7^{\Delta APC}$ and control wild-type littermates treated with DSS. Data are representative of three separate experiments ($n = 5$). $*P < 0.05$, $**P < 0.005$, $***P < 0.0005$. Error bars indicate mean ± s.e.m.

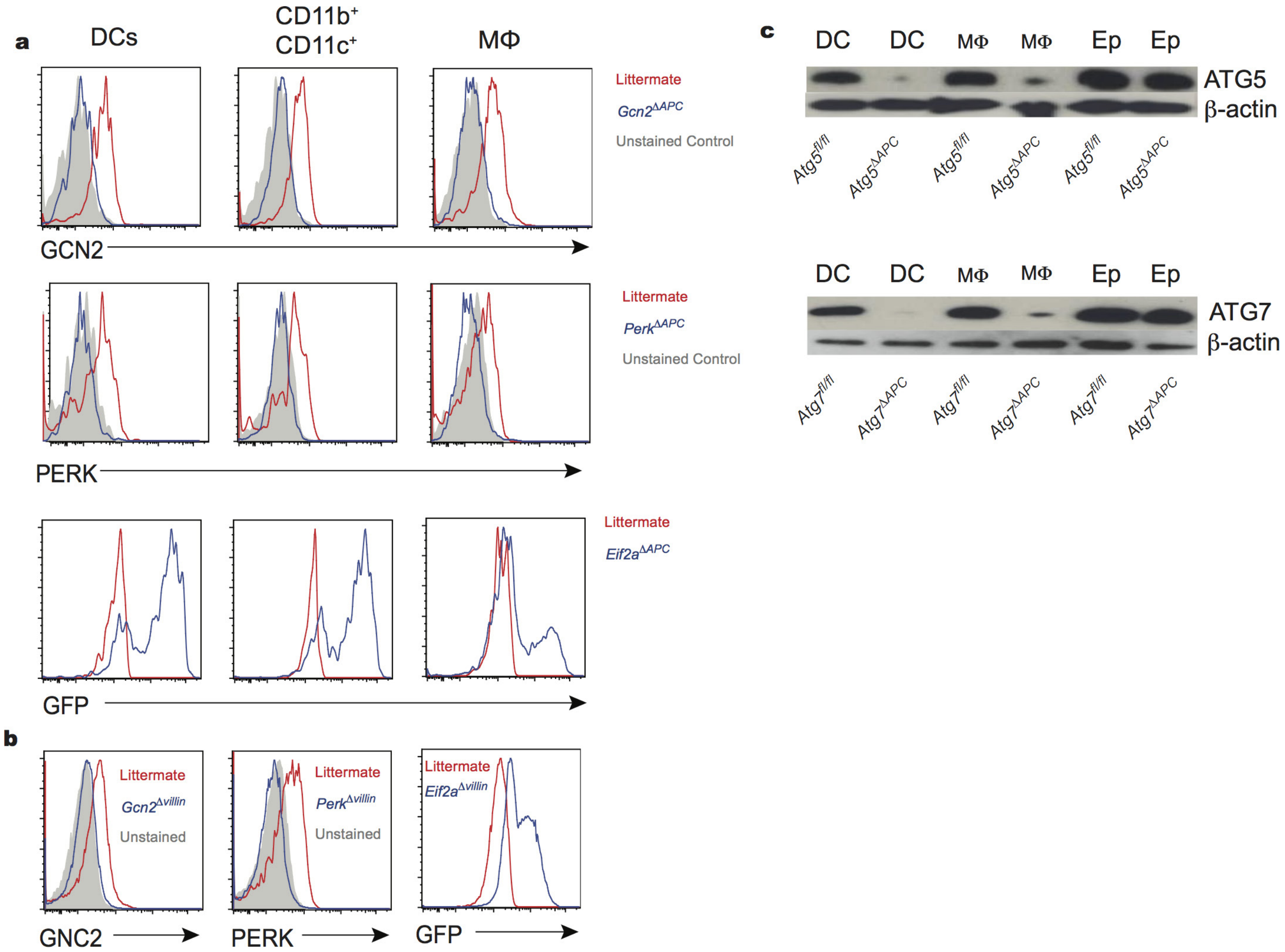


**Extended Data Figure 8 | GCN2-induced autophagy protects the intestinal tissue from the effects of excess oxidation and inflammation.** **a**–**f**, *Gcn2*$^{-/-}$ and littermate wild-type mice were treated with 2% DSS in the drinking water for 5 days. **a**, MFIs of ROS and mitochondrial ROS (MitoSOX) in individual APC subsets and epithelial cells of the small intestine isolated from wild-type and *Gcn2*$^{-/-}$ mice were analysed by flow cytometry ($n = 5$). **b**, MFIs of pro-IL-1β in individual APC subsets and epithelial cells of the small intestine isolated from wild-type or *Gcn2*$^{-/-}$ mice (day 5 after DSS) were analysed by flow cytometry ($n = 3$). **c**, Histology analysis of *Gcn2*$^{-/-}$ mice that were treated with neutralizing antibody. **d**, Effects of low protein diet on DSS-induced colitis. **e**, Colon length. **f**, Frequencies of CD4$^+$, CD4$^+$IFNγ$^+$ and CD4$^+$Foxp3$^+$ T cells. Data are from one experiment that is representative of two or three separate experiments. *$P < 0.05$, **$P < 0.005$, ***$P < 0.0005$. Error bars indicate mean ± s.e.m.




b
Inflammation
Tissue Damage
Tissue Remodeling
Protein Synthesis (Amino acid depletion)
GCN2/eIF2α Activation
Suppression of Inflammasome

**Extended Data Figure 9 | Mechanisms by which GCN2 contributes to protection in the gut against acute colitis. a**, Amino acid starvation induced by an inflamed colon activates GCN2, which triggers autophagy, which is important in inhibiting oxidative stress and pro-IL-1β. Furthermore, levels of IL-1β dictate the magnitude of IL-17A-producing CD4 T cells in the colon. **b**, A hypothetical model for the evolutionary significance of coupling amino acid starvation with control of inflammation.



**Extended Data Figure 10 | Mouse phenotyping by flow cytometry and western blot. a**, **b**, In addition to molecular genotyping via tail DNA, we analysed the protein levels in various mucosal subsets by flow cytometry in various APC-specific (**a**) and epithelial-specific (**b**) conditional knockouts. **c**, Western blot to show selective depletion in APC populations in *Atg5*$^{\Delta\mathrm{APC}}$ and *Atg7*$^{\Delta\mathrm{APC}}$ mice. DC, dendritic cell; Ep, epithelial cell.

doi:10.1038/nature17184

# PGC1α drives NAD biosynthesis linking oxidative metabolism to renal protection

Mei T. Tran[1,2], Zsuzsanna K. Zsengeller[1,2,3], Anders H. Berg[3,4], Eliyahu V. Khankin[1,2], Manoj K. Bhasin[2,5], Wondong Kim[6], Clary B. Clish[7], Isaac E. Stillman[4], S. Ananth Karumanchi[1,2,8], Eugene P. Rhee[6,7] & Samir M. Parikh[1,2]

**The energetic burden of continuously concentrating solutes against gradients along the tubule may render the kidney especially vulnerable to ischaemia. Acute kidney injury (AKI) affects 3% of all hospitalized patients[1,2]. Here we show that the mitochondrial biogenesis regulator, PGC1α[3,4], is a pivotal determinant of renal recovery from injury by regulating nicotinamide adenine dinucleotide (NAD) biosynthesis. Following renal ischaemia, *Pgc1α*$^{-/-}$ (also known as *Ppargc1a*$^{-/-}$) mice develop local deficiency of the NAD precursor niacinamide (NAM, also known as nicotinamide), marked fat accumulation, and failure to re-establish normal function. Notably, exogenous NAM improves local NAD levels, fat accumulation, and renal function in post-ischaemic *Pgc1α*$^{-/-}$ mice. Inducible tubular transgenic mice (iNephPGC1α) recapitulate the effects of NAM supplementation, including more local NAD and less fat accumulation with better renal function after ischaemia. PGC1α coordinately upregulates the enzymes that synthesize NAD *de novo* from amino acids whereas PGC1α deficiency or AKI attenuates the *de novo* pathway. NAM enhances NAD via the enzyme NAMPT and augments production of the fat breakdown product β-hydroxybutyrate, leading to increased production of prostaglandin PGE$_2$ (ref. 5), a secreted autacoid that maintains renal function. NAM treatment reverses established ischaemic AKI and also prevented AKI in an unrelated toxic model. Inhibition of β-hydroxybutyrate signalling or prostaglandin production similarly abolishes PGC1α-dependent renoprotection. Given the importance of mitochondrial health in ageing and the function of metabolically active organs, the results implicate NAM and NAD as key effectors for achieving PGC1α-dependent stress resistance.**

The mature renal tubule returns ~140 l per day of filtered plasma water back to the circulation by establishing energy-intensive electrochemical gradients between the filtrate and vasculature. The kidney is only second to the heart in mitochondrial abundance[6]. We hypothesized that PGC1α (peroxisome proliferator activated receptor gamma co-activator-1-α), enriched in renal tubules and important for stress resistance in the brain, heart and other metabolically active organs[4,7–11], regulates oxidative metabolism in the epithelium to affect overall kidney health.

Hans Krebs identified acylglycerols as a major renal fuel[12]. Following transient local ischaemia, renal function worsened, PGC1α expression declined, tubular mitochondria swelled, and a pronounced accumulation of acylglycerols developed in tubules ($P < 0.0001$, Fig. 1a–e, Extended Data Fig. 1a–c). The fidelity of serum creatinine was confirmed by comparison to cystatin C and inulin clearance (Extended Data Fig. 1d–f). *Pgc1α*$^{-/-}$ mice experienced worse renal function, greater fat accumulation, and more tubular injury following ischaemia (Fig. 1f, Extended Data Fig. 2a–g). To define pathways specific to PGC1α altered by ischaemia, we examined metabolite profiles. Comparing sham with post-ischaemic kidneys yielded six differentially abundant metabolites; comparing uninjured *Pgc1α*$^{-/-}$ to wild-type littermate kidneys yielded 11. Four were shared between settings,

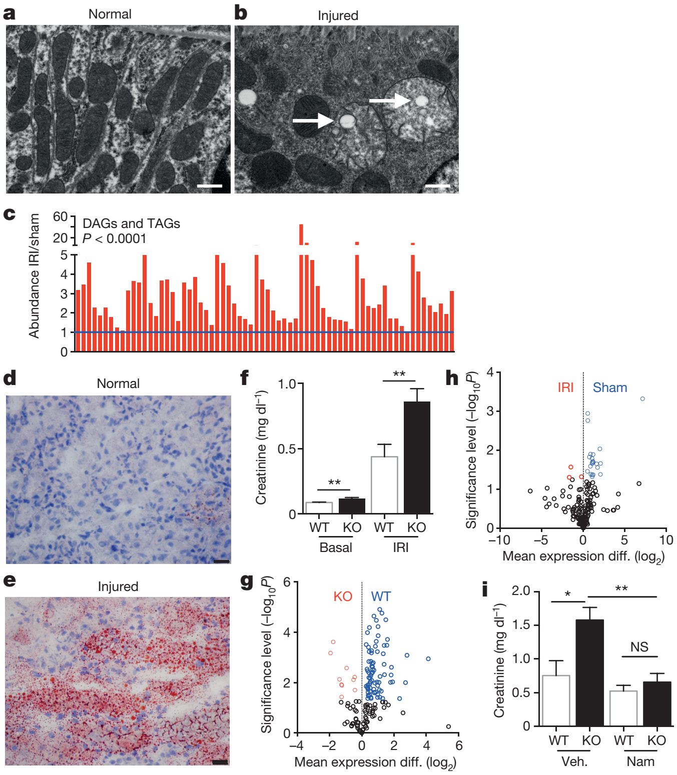


**Figure 1 | NAM supplementation restores normal post-ischaemic response in *Pgc1α*$^{-/-}$ mice. a**, **b**, Pre-ischaemic normal morphology (**a**) and swollen mitochondria (**b**) inside tubular cells 24 h after ischaemia-reperfusion injury (IRI). Scale bars, 200 nm. **c**, Renal di-/tri-acylglycerols (DAGs, TAGs) 24 h following sham or IRI ($n = 6$ per group). *P* value determined by ANOVA. **d**, **e**, Oil-Red-O staining (pink) for fat in normal and post-ischaemic kidneys. Scale bars, 20 μm. **f**, Serum creatinine in wild-type (WT) versus *Pgc1α*$^{-/-}$ (KO) mice (basal, $n = 7$ per group; post-ischaemia, $n = 18$ per group). **g**, **h**, Volcano plots of kidney metabolites from knockout versus wild type or IRI versus sham (univariate $P < 0.05$ for coloured dots, $n = 6$ per group). **i**, Serum creatinine in post-ischaemic wild-type versus knockout mice treated with vehicle (Veh., $n = 5$) versus NAM ($n = 9$). Error bars, s.e.m.; $^{*}P < 0.05$, $^{**}P < 0.01$.

[1] Division of Nephrology and Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA. [2] Center for Vascular Biology Research, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA. [3] Division of Clinical Chemistry, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA. [4] Department of Pathology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA. [5] Bioinformatics and Systems Biology Core, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts 02215, USA. [6] Nephrology and Endocrine Divisions, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts 02114, USA. [7] Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02139, USA. [8] Howard Hughes Medical Institute, Chevy Chase, Maryland 20815, USA.

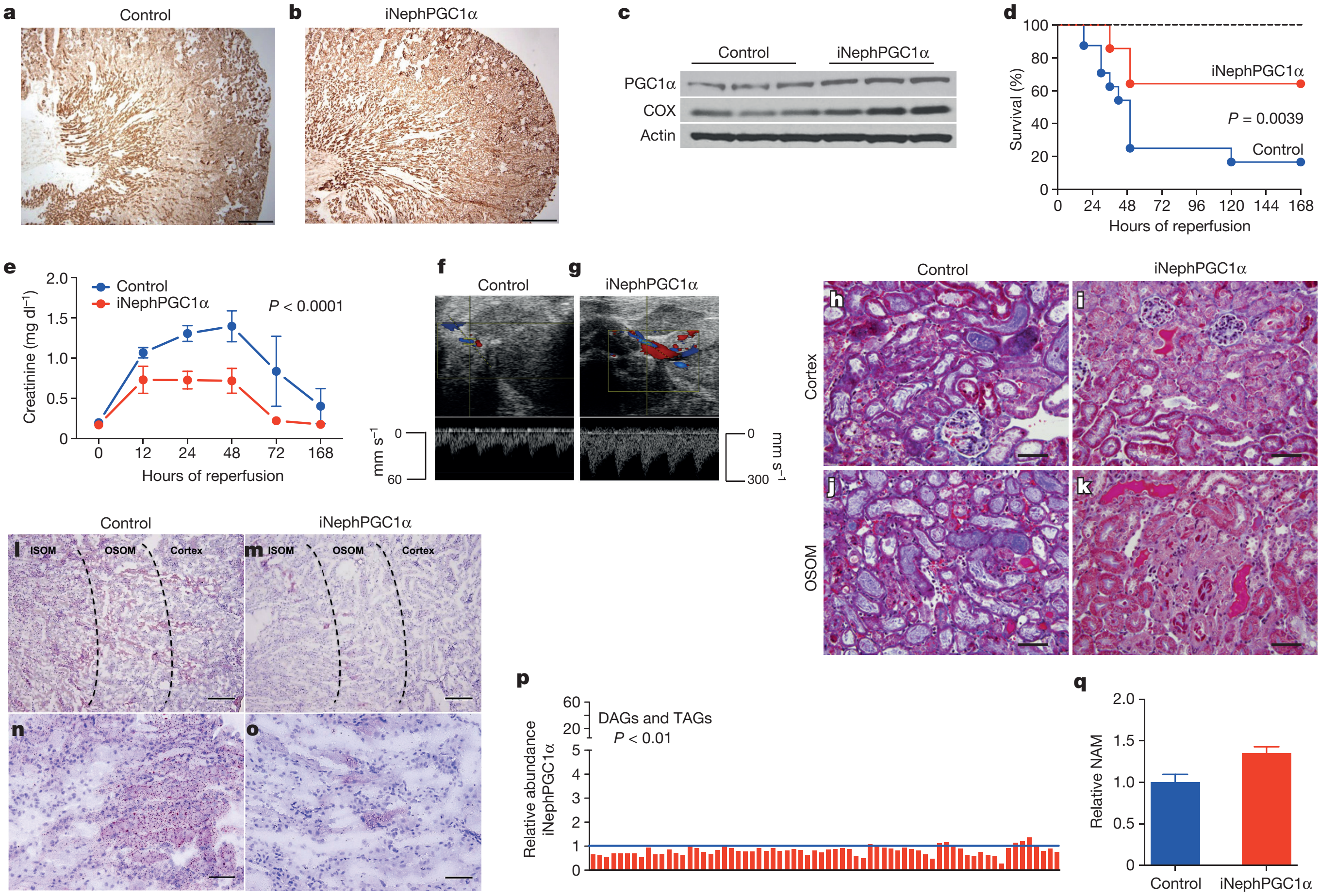


**Figure 2 | Metabolic protection in post-ischaemic iNephPGC1α mice.** **a**, **b**, Renal cytochrome *c* oxidase activity (brown). Scale bar, 500 μm. **c**, Renal PGC1α and cytochrome *c* oxidase subunit IV. **d**, Survival curve following IRI ($n = 14$ control; 24 iNephPGC1α). Dashed line for sham-operated mice ($n = 10$). **e**, Serial serum creatinine levels from mice in **d** analysed by ANOVA. **f**, **g**, Renal artery pulse wave and colour Doppler 24 h after IRI representative of 6 animals per group. **h–k**, Tubular injury in cortex and outer stripe of outer medulla (OSOM) 24 h after IRI representative of 8 animals per group. Scale bar, 100 μm. **l–o**, Oil-red-O staining (pink) in iNephPGC1α mice and controls 24 h after IRI representative of 8 animals per group. Scale bars, 200 μm (top) and 50 μm (bottom). **p**, Renal di-/tri-acylglycerols (DAGs/TAGs) in post-ischaemic iNephPGC1α mice relative to controls ($n = 6$ per group). **q**, Relative renal NAM 24 h after IRI ($n = 6$ per group). Error bars, s.e.m.; $^*P < 0.05$.

with all four lower in *Pgc1α*$^{-/-}$ and post-ischaemic kidneys (Fig. 1g, h, Extended Data Fig. 3a, b).

Of these, carnitine deficiency in *Pgc1α*$^{-/-}$ and post-ischaemic kidneys supported mitochondrial involvement in both situations. Deficiency of betaine and choline, two osmolytes essential for cell volume maintenance in the uniquely hypertonic renal environment, was not unanticipated. We therefore focused on NAM, the predominant mammalian precursor to synthesize the energy carrier NAD needed for fatty acid oxidation (FAO)[13]. After confirming the metabolomics results (Extended Data Fig. 3c–e), we tested the effect of NAM supplementation. Exogenous NAM increased renal NAM ($P < 0.001$), normalized post-ischaemic fat accumulation, and completely prevented post-ischaemic AKI in *Pgc1α*$^{-/-}$ mice (Fig. 1i, Extended Data Fig. 3f–h), implicating this metabolite as an unexpected effector of PGC1α.

To probe the robustness of PGC1α's relation to NAM, fat accumulation, and renal function, we developed an inducible tubular epithelial transgenic model using the well-validated Pax8 promoter (iNephPGC1α)[14]. Heterologous PGC1α was tightly controlled without leaky gene expression; organ size and mass were indistinguishable; and mitochondrial abundance increased—as assessed by comparing mitochondrial to nuclear DNA and mitochondrial gene products to cytosolic gene products—without altering ultrastructural morphology or the anatomical distribution favouring cortex and outer stripe of the outer medulla (Fig. 2a–c, Extended Data Fig. 4a–i). iNephPGC1α mice tolerated renal ischaemia more successfully, achieving better survival ($P = 0.0039$), more preserved function ($P < 0.0001$), better kidney perfusion, and less tubular injury (Fig. 2d–k, Extended Data Fig. 4j, k). Sham-operated mice experienced no significant change in creatinine or reduced survival. Renal NAM was higher in post-ischaemic iNephPGC1α mice, and post-ischaemic fat accumulation was markedly reduced compared to controls ($P < 0.01$, Fig. 2l–q). Renal protection in iNephPGC1α mice was shared across distinct models as post-inflammatory renal injury was also attenuated (Extended Data Fig. 5a). PGC1α's effect appeared to be cell-type specific as endothelial overexpression conferred no renoprotection (Extended Data Fig. 5b).

RNA sequencing identified 1,160 transcripts associated with PGC1α-dependent renoprotection (Fig. 3a, Supplementary Information Table 1). The pathways most over-represented related to intermediary metabolism (Fig. 3b). Closer examination revealed that *de novo* NAD biosynthetic enzymes were coordinately regulated, induced in uninjured iNephPGC1α kidneys and suppressed in post-ischaemic or uninjured *Pgc1α*$^{-/-}$ kidneys (Fig. 3c–f). The effect of PGC1α on the *de novo* pathway was cell-autonomous, as knockdown in isolated renal tubular cells was sufficient to suppress the pathway ($P = 0.0001$, Extended Data Fig. 6a).

As epithelial PGC1α defended renal function and resolved post-ischaemic fat accumulation, we hypothesized that protection from AKI may relate to NAM, NAD, and fatty acid utilization. Indeed, exogenous

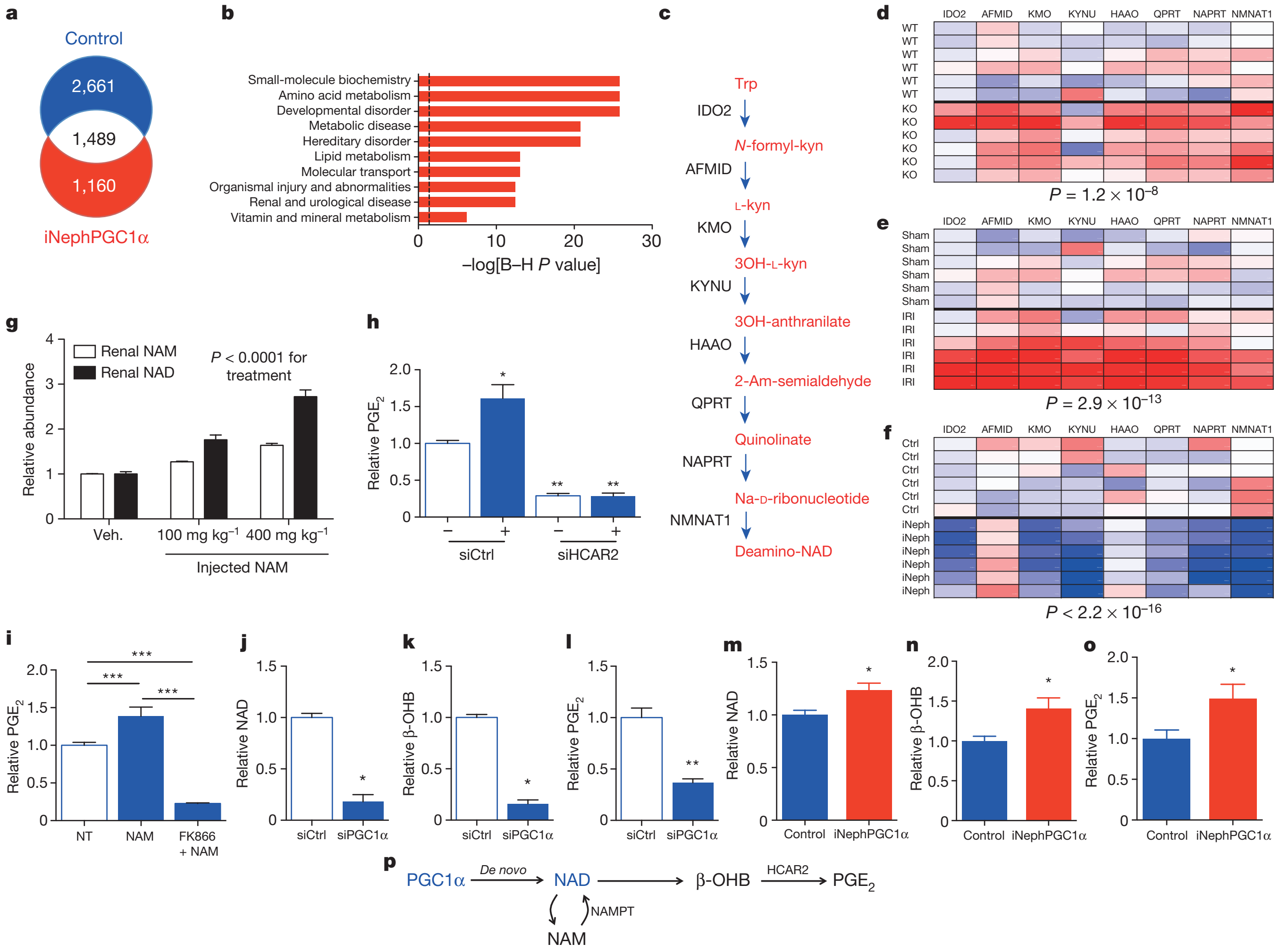


**Figure 3 | NAM induces β-OHB downstream of PGC1α to augment PGE$_2$.** **a**, Renal RNA sequencing 24 h after IRI or sham operation in controls versus iNephPGC1α mice with enumerated transcripts. **b**, Pathway analysis of 1,160 transcripts unique to post-ischaemic iNephPGC1α mice graphed by $-\log_{10}$[Benjamini–Hochberg-corrected *P* value]. Dashed line at $P < 0.05$. **c**, *De novo* NAD biosynthetic pathway adapted from KEGG (http://www.genome.jp/kegg). Trp, tryptophan, kyn, kynurenine, Am, amino, Na, nicotinate. **d**–**f**, Heat maps of intrarenal expression (red, lower; blue, higher) for *de novo* pathway from knockout (KO) versus wild type (WT); 24 h after sham versus IRI; and iNephPGC1α versus controls ($n = 6$ per group). *P* values by ANOVA. **g**, Relative renal NAM and NAD 4 h after indicated NAM dose. *P* value by ANOVA. **h**, Conditioned-media-PGE$_2$ of renal tubular cells after HCAR2 knockdown with and without HCAR2 stimulation (+, niacin 10 mM, $n = 6$ per condition). **i**, PGE$_2$ from renal cells following NAM (1 μM for 24 h) with and without NAMPT inhibitor FK866 (10 nM, $n = 6$ per condition). **j**–**l**, Intracellular NAD, conditioned-media β-hydroxybutyrate (β-OHB), and conditioned-media PGE$_2$ in PGC1α knockdown cells ($n = 6$ per condition). **m**–**o**, Relative renal NAD, β-OHB, and PGE$_2$ in control versus iNephPGC1α mice ($n = 6$ per group). $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. **p**, Renal epithelial PGC1α coordinately upregulates *de novo* NAD biosynthesis, in the absence of which NAM is used through the NAMPT-salvage pathway to generate NAD. Consequently, β-OHB accumulates, which signals HCAR2 to induce PGE$_2$. Error bars, s.e.m.

NAM dose-dependently increased renal NAD and drove local accumulation of the fatty acid breakdown product β-hydroxybutyrate (β-OHB) to approximately tenfold higher than normal circulating concentrations ($P < 0.0001$, Fig. 3g and Extended Data Fig. 6b, c). β-OHB activates HCAR2, a G-protein coupled receptor that induces the renoprotective prostaglandin PGE$_2$ (ref. 5). Silencing or chemical inhibition of HCAR2 markedly reduced both basal and ligand-dependent PGE$_2$ secretion (Fig. 3h, Extended Data Fig. 6d, e). NAM augmented PGE$_2$ secretion, requiring conversion to NAD via the enzyme NAMPT to do so (Fig. 3i, Extended Data Fig. 6f, g)[15]. Silencing of PGC1α reduced each intermediate, lowering the cellular NAD and secreted β-OHB and PGE$_2$ (Fig. 3j–l). In PGC1α-silenced cells, excess β-OHB was still able to induce PGE$_2$ secretion ($P < 0.0001$, Extended Data Fig. 6h). Finally, renal levels of each component mirrored the cellular results, with opposing effects of PGC1α deficiency and excess on NAD, β-OHB, and PGE$_2$ (Fig. 3m–o, Extended Data Fig. 7a–c). Together, these results implicated PGC1α-dependent NAD production as an important determinant of cellular metabolism that induces renoprotective molecules (Fig. 3p).

To test this further, we inhibited β-OHB signalling with mepenzolate bromide or prostaglandin synthesis with indomethacin in iNephPGC1α mice subjected to ischaemia. Renal protection was similarly abolished in either setting, confirming their roles as PGC1α effectors (Fig. 4a, b, Extended Data Fig. 7d, e). Since NAM prevented ischaemic AKI in *Pgc1α*$^{-/-}$ mice, we then asked whether NAM has a broader therapeutic role. NAM administered after established AKI significantly improved renal function ($P = 0.0011$, Fig. 4c). We also observed that renal NAM declined following cisplatin, a chemotherapy that injures the kidney through a mechanism considered distinct from ischaemia (Extended Data Fig. 7f, g). NAM supplementation prevented cisplatin-induced AKI (Fig. 4d, e). Finally, we found that PGC1α expression in human AKI was strongly suppressed, even in histologically normal regions of renal tissue (Fig. 4f–h, Extended Data Fig. 8a–f), mirroring the AKI-induced suppression of PGC1α observed in experimental models (Extended Data Fig. 1c and ref. 7). These results show that PGC1α is a negatively regulated target in AKI.

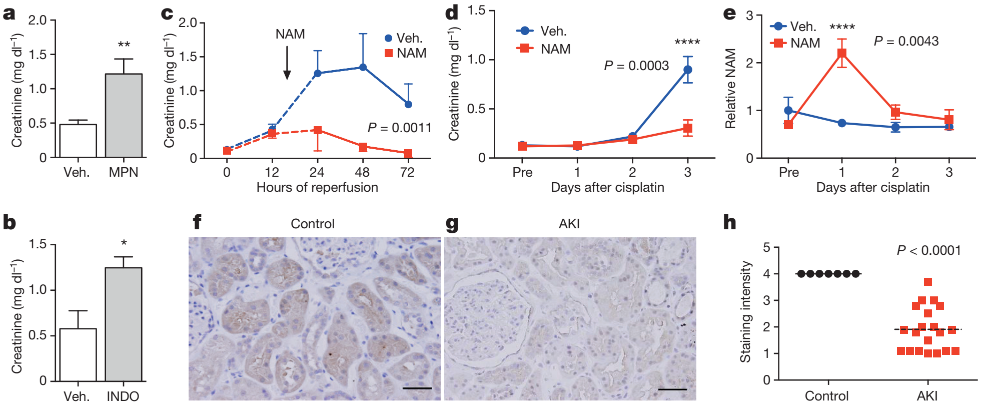


**Figure 4 | PGC1α effectors, NAM as therapy, and PGC1α in human AKI. a**, Serum creatinine in iNephPGC1α mice 24 h after IRI with vehicle versus mepenzolate (MPN, 10 mg kg$^{-1}$ intraperitoneal) treatment ($n = 6$ per group). **b**, Serum creatinine in iNephPGC1α mice 24 h after IRI with vehicle versus indomethacin (INDO, 10 mg kg$^{-1}$ intraperitoneal) treatment ($n = 6$ per group). **c**, Serial serum creatinines in mice receiving a single dose of NAM (400 mg kg$^{-1}$ intraperitoneal) 18 h after the onset of reperfusion, that is, with established AKI. Analysed by ANOVA ($n = 5$ per group). **d**, Serial serum creatinines after cisplatin (25 mg kg$^{-1}$ intraperitoneal administered on day 0) with or without NAM (400 mg kg$^{-1}$ intraperitoneal on day −1 and day 0). Analysed with Bonferroni-corrected ANOVA ($n = 5$ per group). **e**, Relative renal NAM from **d**. **f**, **g**, Representative immunostaining (brown) for PGC1α from control human kidney and a renal biopsy for AKI. Scale bars, 50 μm. **h**, PGC1α immunostaining intensity (1, weakest; 4, strongest). Each dot represents a unique specimen. Analysed by Mann–Whitney *U*-test. Error bars, s.e.m.; $^*P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$.

The present results identify PGC1α as a pivotal mediator of renal resistance to acute stressors. By linking oxidative metabolism in the epithelium to overall organ function, the proposed pathway provides new insight into an old observation, namely the exquisite sensitivity of the kidney to ischaemia and other insults. More fundamentally, the results implicate NAD biosynthesis as a coordinately regulated target of PGC1α, a mechanism that may contribute to the PGC1α-dependent stress resistance previously observed across several metabolically active organs[7–11].

NAD has long been recognized for its central role in energy metabolism, with recent work demonstrating that NAD is rate-limiting for mitochondrial function[16]. NAD augmentation appears to restore youthful mitochondrial function and reverse age-related declines in health[17]. In contrast, NAD depletion has been described as a feature of diabetes[18]. Since diabetes and ageing are two of the most prevalent predispositions for AKI, the present results motivate interest in whether local NAD concentration may provide a set point for resistance to acute renal stressors. NAD may also be important for the gradual decline of kidney function with normal ageing.

That an even larger set of known AKI risk factors—including diabetes, but also chronic kidney disease (Extended Data Fig. 8g, h), sepsis, and warm ischaemia—is associated with reduction of PGC1α[7] further attests to the potential relevance of the results to human disease. Experiments targeting mitochondrial biogenesis through a drug-screening approach offer additional promise for this avenue in AKI[19]. Since AKI has been associated with death in critically ill patients[1], that excess renal PGC1α improves survival after AKI highlights the importance of the kidney to overall health. Downstream of PGC1α, NAM may not only be an effective preventative agent, but also a potential therapy for established AKI, a set of diseases for which no drug has yet been identified.

PGC1α in skeletal muscle has been shown to exert extracellular effects, whether through the myokine irisin, metabolites such as kynurenine, or the angiogenic factor VEGF[8,9,20]. By comparison, the present results show that renal tubular PGC1α communicates with neighbouring cells at least through $PGE_2$. Therapeutic manipulation of renal β-OHB may constitute one means of increasing $PGE_2$. $PGE_2$ is a well-recognized vasodilator in the kidney, but may also be exerting cytoprotective effects in AKI (reviewed in ref. 21), actions that have been demonstrated in multiple animal models and even humans[22–24].

We observed enhanced renal function, vascular relaxation, and increased perfusion at baseline as a result of excess PGC1α in the epithelial compartment of the kidney in the iNephPGC1α model (Extended Data Fig. 9a–j), physiological features that would be consistent with functional responses of the local vasculature to the excess renal $PGE_2$ present in this model. However, we also asked whether VEGF was regulated by renal PGC1α as such features could also arise from increased vascularization (Extended Data Fig. 9k–m). PGC1α$^{-/-}$ mice showed no decrement in renal VEGF and iNephPGC1α mice displayed only modest induction versus their respective controls. This strongly contrasts with VEGF induction by skeletal muscle PGC1α[8], suggesting the presence of cell-specific modulators of PGC1α function such as ERRα, which is notably more abundant in skeletal muscle than kidney (http://www.biogps.org).

Our results suggest several avenues for future investigation. First, the coordinated regulation of NAD biosynthesis by PGC1α may occur in other cells and organs, particularly under stress conditions. Of the major biosynthetic routes to NAD—*de novo* from amino acids, the Preiss–Handler pathway from niacin, and salvage from NAM via NAMPT—the action of PGC1α on each will require careful dissection as this may vary depending on cell type and condition. The transcription factor(s) with which PGC1α interacts to induce the *de novo* pathway are of substantial interest. Since ischaemia did not reduce renal NAMPT (Extended Data Fig. 10), the salvage pathway may be a viable therapeutic route. Second, the rapid reduction of NAD during AKI may also relate to its already short half-life[25] as well as the action of NAD-consuming enzymes such as PARPs, nucleotidases, and sirtuins, all of which have been implicated in this condition[26–28]. Third, NAD's emerging role as a guardian against age-related decline in health and mitochondrial function[17] suggests that therapeutic manipulation of NAM and NAD may have implications beyond AKI. For example, NAMPT agonism protects against experimental neuronal injury[29], and inhibition of urinary NAM disposal (by blocking N-methylation) prevents experimental obesity[30]. Finally, the link from mitochondrial metabolism to renoprotective prostaglandins unites two major avenues of mechanistic investigation in AKI, but other mediators and downstream effectors for renal PGC1α may also exist.

In summary, the present work applies complementary discovery approaches to identify a new pathway by which parenchymal PGC1α affects NAD to protect against renal injury. The results may also have mechanistic, diagnostic, and therapeutic implications for ischaemia in the heart, brain and other metabolically active organs.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 14 May 2015; accepted 27 January 2016.**
**Published online 16 March 2016.**

1. Thadhani, R., Pascual, M. & Bonventre, J. V. Acute renal failure. *N. Engl. J. Med.* **334,** 1448–1460 (1996).
2. Lewington, A. J., Cerda, J. & Mehta, R. L. Raising awareness of acute kidney injury: a global perspective of a silent killer. *Kidney Int.* **84,** 457–467 (2013).
3. Puigserver, P. *et al.* A cold-inducible coactivator of nuclear receptors linked to adaptive thermogenesis. *Cell* **92,** 829–839 (1998).
4. Ruas, J. L. *et al.* A PGC-1$\alpha$ isoform induced by resistance training regulates skeletal muscle hypertrophy. *Cell* **151,** 1319–1331 (2012).
5. Hanson, J. *et al.* Nicotinic acid- and monomethyl fumarate-induced flushing involves GPR109A expressed by keratinocytes and COX-2-dependent prostanoid formation in mice. *J. Clin. Invest.* **120,** 2910–2919 (2010).
6. Pagliarini, D. J. *et al.* A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134,** 112–123 (2008).
7. Tran, M. *et al.* PGC-1$\alpha$ promotes recovery after acute kidney injury during systemic inflammation in mice. *J. Clin. Invest.* **121,** 4003–4014 (2011).
8. Arany, Z. *et al.* HIF-independent regulation of VEGF and angiogenesis by the transcriptional coactivator PGC-1$\alpha$. *Nature* **451,** 1008–1012 (2008).
9. Agudelo, L. Z. *et al.* Skeletal muscle PGC-1$\alpha$1 modulates kynurenine metabolism and mediates resilience to stress-induced depression. *Cell* **159,** 33–45 (2014).
10. Arany, Z. *et al.* Transverse aortic constriction leads to accelerated heart failure in micelacking PPAR-$\gamma$ coactivator 1$\alpha$. *Proc. Natl Acad. Sci. USA* **103,** 10086–10091 (2006).
11. St-Pierre, J. *et al.* Suppression of reactive oxygen species and neurodegeneration by the PGC-1 transcriptional coactivators. *Cell* **127,** 397–408 (2006).
12. Weidemann, M. J. & Krebs, H. A. The fuel of respiration of rat kidney cortex. *Biochem. J.* **112,** 149–166 (1969).
13. Collins, P. B. & Chaykin, S. The management of nicotinamide and nicotinic acid in the mouse. *J. Biol. Chem.* **247,** 778–783 (1972).
14. Traykova-Brauch, M. *et al.* An efficient and versatile system for acute and chronic modulation of renal tubular function in transgenic mice. *Nature Med.* **14,** 979–984 (2008).
15. Revollo, J. R., Grimm, A. A. & Imai, S. The NAD biosynthesis pathway mediated by nicotinamide phosphoribosyltransferase regulates Sir2 activity in mammalian cells. *J. Biol. Chem.* **279,** 50754–50763 (2004).
16. Bai, P. *et al.* PARP-1 inhibition increases mitochondrial metabolism through SIRT1 activation. *Cell Metab.* **13,** 461–468 (2011).
17. Gomes, A. P. *et al.* Declining NAD$^+$ induces a pseudohypoxic state disrupting nuclear-mitochondrial communication during aging. *Cell* **155,** 1624–1638 (2013).
18. Garcia Soriano, F. *et al.* Diabetic endothelial dysfunction: the role of poly(ADP-ribose) polymerase activation. *Nature Med.* **7,** 108–113 (2001).
19. Jesinkey, S. R. *et al.* Formoterol restores mitochondrial and renal function after ischemia-reperfusion injury. *J. Am. Soc. Nephrol.* **25,** 1157–1162 (2014).
20. Boström, P. *et al.* A PGC1-$\alpha$-dependent myokine that drives brown-fat-like development of white fat and thermogenesis. *Nature* **481,** 463–468 (2012).
21. Breyer, M. D., Jacobson, H. R. & Breyer, R. M. Functional and molecular aspects of renal prostaglandin receptors. *J. Am. Soc. Nephrol.* **7,** 8–17 (1996).
22. Papanicolaou, N., Callard, P., Bariety, J. & Milliez, P. The effect of indomethacin and prostaglandin (PGE2) on renal failure due to glycerol in saline-loaded rats. *Clin. Sci. Mol. Med.* **49,** 507–510 (1975).
23. Mauk, R. H., Patak, R. V., Fadem, S. Z., Lifschitz, M. D. & Stein, J. H. Effect of prostaglandin E administration in a nephrotoxic and a vasoconstrictor model of acute renal failure. *Kidney Int.* **12,** 122–130 (1977).
24. Sketch, M. H., Jr *et al.* Prevention of contrast media-induced renal dysfunction with prostaglandin E1: a randomized, double-blind, placebo-controlled study. *Am. J. Ther.* **8,** 155–162 (2001).
25. Feldkamp, T. *et al.* Preservation of complex I function during hypoxia-reoxygenation-induced mitochondrial injury in proximal tubules. *Am. J. Physiol. Renal Physiol.* **286,** F749–F759 (2004).
26. Morigi, M. *et al.* Sirtuin 3-dependent mitochondrial dynamic improvements protect against acute kidney injury. *J. Clin. Invest.* **125,** 715–726 (2015).
27. Ebrahimkhani, M. R. *et al.* Aag-initiated base excision repair promotes ischemia reperfusion injury in liver, brain, and kidney. *Proc. Natl Acad. Sci. USA* **111,** E4878–E4886 (2014).
28. Rajakumar, S. V. *et al.* Deficiency or inhibition of CD73 protects in mild kidney ischemia-reperfusion injury. *Transplantation* **90,** 1260–1264 (2010).
29. Wang, G. *et al.* P7C3 neuroprotective chemicals function by activating therate-limiting enzyme in NAD salvage. *Cell* **158,** 1324–1334 (2014).
30. Kraus, D. *et al.* Nicotinamide *N*-methyltransferase knockdown protects against diet-induced obesity. *Nature* **508,** 258–262 (2014).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** The authors thank B. Spiegelman (Harvard Medical School, Dana Farber Cancer Institute) for support from this project's inception; M. Zeidel (Harvard Medical School, BIDMC) for advice; Z. Arany (University of Pennsylvania) for the VE-cadherin-tTA × TRE-PGC1$\alpha$ mice; A. Agarwal (University of Alabama, NIH P30-DK079337) for LC–MS measurements of serum creatinine; A. Kurmann and A. Hollenberg (Harvard Medical School, BIDMC) for thyroxine measurements; and P. Pacher (NIH) for cisplatin-treated kidneys for microscopy. This work was supported by R01-DK095072 and philanthropic funds to S.M.P.; K08-DK090142 and a grant from Satellite Healthcare to E.P.R.; and K08-DK101560 to E.V.K.

**Author Contributions** M.T.T. designed experiments, performed the breeding, genotyping, renal injury models, cellular studies, analysed data, and wrote the manuscript. Z.K.Z. and I.E.S. performed and analysed histopathology, enzyme histochemistry, electron microscopy, and the human biopsy immunohistochemistry studies. A.H.B. created LC–MS assays, measured metabolites for cellular experiments, and analysed metabolic results. E.V.K. and S.A.K. performed micro-ultrasounds and analysed flow results. M.K.B. analysed raw RNA sequencing data. W.K., C.B.C. and E.P.R. performed metabolomics, follow-up metabolite measurements, and *in vivo* experiments with cisplatin and NAM. S.M.P. designed the experiments, analysed results, and wrote the manuscript with input from all authors.

**Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to S.M.P. (sparikh1@bidmc.harvard.edu).

## METHODS

**Mouse studies.** All studies with mice were reviewed and approved by the Institutional Animal Use and Care Committee of Beth Israel Deaconess Medical Center (BIDMC). *Pgc1*$\alpha^{-/-}$ (stock no. 008597), Pax8-rtTA (no. 007176) and TRE-PGC1α (no. 012387) mice were all obtained from Jackson Laboratories and bred at BIDMC. The parental strains were generated on a mixed C57 background with further backcrossing into C57BL/6J as described by the manufacturer, except for the TRE-PGC1α mouse, which was generated on and is maintained on FVB. Primers for genotyping have been described elsewhere[31,32]. All experiments were performed with genetically appropriate littermate controls.

Ischaemia-reperfusion injury (IRI) was performed on 8–12-week-old males through two small paramedial dorsal incisions by applying a microvascular clamp to each renal pedicle for 20 min. Mice were anaesthetized with isoflurane for the duration of surgery and warmed to 37 °C using a servo-controlled heating pad. Incisions were closed in two layers and mice were revived with 1 ml warm saline injected intraperitoneally.

All chemicals were purchased from Sigma-Aldrich unless otherwise noted. NAM was given by intraperitoneal injection of 400 mg$^{-1}$ kg$^{-1}$ day$^{-1}$ for 4 days in saline, with the final dose an hour before IRI surgery. In rescue experiments, the same dose was administered once 18 h after reperfusion. Indomethacin was given by intraperitoneal injection of 10 mg kg$^{-1}$ in 0.1 M sodium carbonate/saline an hour before IRI. The HCAR2 inhibitor, mepenzolate bromide, was given by intraperitoneal injection of 10 mg kg$^{-1}$ in saline an hour before IRI[33-35]. LPS (*E.coli* serotype O111:B4) was given by intraperitoneal injection of 25 mg kg$^{-1}$ in saline. Cisplatin was given by intraperitoneal injection of 25 mg kg$^{-1}$ as previously described[36]. Unless otherwise stated, blood and kidneys were collected 24 h after the AKI model. The experiments were not randomized.

**Mass spectrometry measurements.** All measurements were performed in a blinded fashion by an independent facility. Creatinine was analysed by LC/MS-MS at the University of Alabama Birmingham O'Brien Core Center for Acute Kidney Injury Research (NIH P30-DK079337). This method adds the accuracy of MS to the LC method of creatinine measurement endorsed by a renal investigative consortium (diacomp.org). The coefficient of variation was 6% indicating high assay precision.

For metabolomics measurements, snap frozen kidneys were cut to equal weights (20 mg per specimen) and mechanically homogenized into four volumes of ice-cold water. Metabolites were assayed as previously described[37]. In brief, amino acids, amines, acylcarnitines, nucleotides, and other cationic polar metabolites were measured in 10 μl of kidney homogenate using hydrophilic interaction liquid chromatography coupled with non-targeted, positive ion mode MS analysis on an Exactive Plus Orbitrap MS (Thermo Scientific). Polar and non-polar lipids were measured in 10 μl of kidney homogenate using C8 chromatography and non-targeted, positive ion mode MS analysis on a Q Exactive MS (Thermo Scientific). Identification of known metabolites was achieved by matching retention times and mass-to-charge ratio ($m/z$) to synthetic mixtures of reference compounds and characterized pooled plasma reference samples. Results were analysed in MetaboAnalyst (http://www.metaboanalyst.ca).

LC–MS assays were developed for multiplex quantification of NAM, NAD, and β-OHB from cellular experiments. NAD measurements reflect total NAD$^+$ plus NADH. In brief, conditioned medium was extracted with methanol (80% methanol final concentration) spiked with isotopic standards for NAM and β-OHB (CDN Isotopes, Inc.). Precipitated proteins were removed by centrifugation, and supernatants were analysed directly. For analysis of cell lysates, cells were washed with ice-cold PBS, scraped and lysed on dry ice into methanol containing isotopic standards. After extraction, cell and media supernatants were analysed by LC–MS/MS using reverse-phase chromatography (NAM and NAD/NADH) or hydrophilic interaction chromatography (β-OHB) coupled to tandem mass spectrometry using an API 5000 triple quadruple mass spectrometer. Analytes were quantified by multiple reaction monitoring using the following $m/z$ transitions: β-OHB 103.1 > 59, β-OHB IS 105.1 > 60, NAM 123.3 > 80.2, NAM IS 127.3 > 84.2, NAD/NADH 664.2 > 542.0. Eluting peaks were quantified by area under the curve (AUC).

Raw AUC values were divided by the mean value of the control group for each experiment, thus the results are presented as relative concentrations to the control group. All assays were performed in triplicate and replicate measurements demonstrated a CV < 5%.

**RNA-seq sequencing and identification of differentially expressed transcripts.** Poly(A)-enriched RNA was isolated from whole kidneys and checked for quality by denaturing agarose gel as well as an Agilent Bioanalyzer. Sequencing libraries were generated from the double-stranded cDNA using the Illumina TruSeq kit according to the manufacturer's protocol. Library quality control was checked using the Agilent DNA High Sensitivity Chip and qRT–PCR. High quality libraries were sequenced on an Illumina HiSeq 2000. To achieve comprehensive coverage for each sample, we generated ~25–30 million single-end reads. Raw results were passed through quality controls steps and alig ned to the mouse genome. Gene expression measurement was performed from aligned reads by counting the unique reads. The read count based gene expression data was normalized on the basis of library complexity and gene variation. The normalized count data was compared among groups using a negative binomial model to identify differentially expressed genes. The differentially expressed genes were identified on the basis of raw $P$ value and fold change. Genes were considered significantly differentially expressed if the multiple test corrected $P$ value was <0.05 and absolute fold change >2.

**Functional enrichment analysis.** Ingenuity Pathway Analysis (IPA 8.0, Qiagen) was used to identify the functions that are significantly affected by significantly differentially expressed genes from different comparisons. The knowledge base of this software consists of functions, pathways, and network models derived by systematically exploring the peer reviewed scientific literature. A detailed description of IPA analysis is available at the Ingenuity Systems' website (http://www.ingenuity.com). A $P$ value is calculated for each function according to the fit of the user's data to the IPA database using one-tailed Fisher exact test. The functions with multiple-test-corrected $P$ values <0.01 were considered significantly affected.

**Western analysis.** Kidney lysate preparation, gel electrophoresis, transfer, immunoblotting, detection, and image acquisition were performed as previously described[31]. Antibodies against PGC1α (Cayman Chemical), cytochrome *c* oxidase subunit IV (Cell Signaling Technology), and Transcription Factor A Mitochondrial, TFAM (Abcam) were used as previously described[31,38].

**Quantitative PCR.** Total RNA extraction and cDNA synthesis were performed as previously described[31]. PCR reactions were performed in duplicate using the ABI 7500 Fast Real-Time PCR and TaqMan gene expression assays (Applied Biosystems). The following TaqMan gene probes were used: *Ppargc1a*, *Ndufs1*, *Cycs*, *Atp5o*, *Nrf1*, *Tfam*, *Vegfa*, *Nos1*, *Nos3* and *Hcar2*. Of the four known *Ppargc1a* transcripts (1–4), *Ppargc1a1* (Taqman Mm00447183_m1) was studied in all gene expression analyses[39]. Mouse *Ido2*, *Afmid*, *Kynu*, *Kmo*, *Haao*, *Qprt*, *Naprt* and *Nmnat1* for SYBR Green PCR have been described elsewhere[40,41]. Mouse *Nampt* SYBR primers were designed using PrimerQuest Tool (Integrated DNA Technologies). Relative expression levels were determined using the comparative threshold method.

**Mitochondrial DNA copy number analysis.** Total DNA was extracted from mouse kidneys using the DNeasy Blood and Tissue Kit (Qiagen) with on-column RNase digestion per manufacturer's instructions. Gene expression of mitochondrial-encoded NADH dehydrogenase 1 (*mt-Nd1*) relative to nuclear 18S rRNA was used to determine mitochondrial DNA copy number as previously described[42].

**Histopathology.** Formalin-fixed, paraffin-embedded blocks were sectioned and stained with H & E, PAS, and Masson trichrome. Ten random high-power fields in the cortex and ten random high-power fields in the outer stripe of the outer medulla were viewed and graded for tubular necrosis—defined as the loss of the proximal tubular brush border, blebbing of apical membranes, tubular necrosis/apoptosis and epithelial cell detachment from the basement membrane or intraluminal aggregation of necrotic debris. Each high-power field was separately scored on a scale (0, no necrosis; 1, rare single necrotic cells; 2, frequent single necrotic cells; 3, groups of necrotic cells; and 4, confluent tubular necrosis) and the average score was compiled for each specimen and then used for between-group comparisons. All scoring was performed by a single operator blinded to genotype and experimental model (IES).

***In situ* COX enzyme chemistry.** Enzyme histochemistry to detect cytochrome *c* oxidase (COX) activity was performed on 6-μm snap-frozen sagittal sections as previously described[31]. Functional electron microscopy used in the cisplatin kidney injury model was described earlier[36].

**Electron microscopy.** The complete method is previously described[31]. In brief, kidneys were fixed with 1.25% glutaraldehyde in 0.1 M cacodylate buffer (pH 7.4) and cut into 1-μm sections in both sagittal and transverse planes for image analysis. After drying the sections, slides were stained at 65 °C for 20 min in 0.1% Toluidine blue in 1% sodium borate, cooled to room temperature, washed in distilled water, cleaned in xylene, and mounted in Permount sections for light microscopy. Subsequent ultrathin sections (0.5 μm) were examined by transmission electron microscopy (JEOL 1011, JEOL Corp.) with Orca-HR Digital Camera (Hamamatsu Corp.), and Advanced Microscopy Technique Corporation image capture system.

**Oil-Red-O staining.** Oil-Red-O solution was prepared by dissolving 0.5 g Oil-Red-O (Poly Scientific) in 100 ml isopropanol. Frozen sections were cut to 5 μm and natively stained in Oil-Red-O solution for 20 min at room temperature, then rinsed in running tap water for 2 min. Haematoxylin counter-staining was performed without differentiation in HCl–ethanol and sections were rinsed with water, then mounted with VectaMount AQ Aqueous Mounting medium (Vector Labs).

**Human biopsy series.** All studies were approved by the Institutional Review Board at BIDMC. Control specimens came from normal tissue sections of nephrectomies. CKD diagnoses included focal segmental glomerulosclerosis, chronic allograft nephropathy, chronic interstitial nephritis, and chronic IgA nephropathy. AKI diagnoses included acute ischaemic injury, post-transplant delayed graft function attributable to ischaemia-reperfusion injury, and acute interstitial nephritis. PGC1α antibody (Abcam ab54481) was used at a dilution of 1:100 and developed with horseradish peroxidase (ImmPRESS polymer staining kit, Vector Labs). The peptide immunogen SKYDSLDFDSLLKEAQRSLRR (synthesized by the Biopolymers Lab, Koch Institute at MIT) was pre-incubated in 100-fold excess of the PGC1α antibody to confirm antibody specificity in human IHC studies. Ten randomly selected high-powered fields were viewed per specimen, with each field scored on a 4-point scale (1, weakest; 4, strongest) based on the intensity of staining, specifically in non-necrotic areas and unscarred areas and avoiding obvious collecting ducts. The average score of each specimen was then used for between-group comparisons. All scoring was performed by a single operator blinded to the underlying diagnosis (IES).

**Micro-ultrasound.** The full method is previously described[31]. In brief, mice were lightly anaesthetized, secured to a heat-controlled stage, and continuously monitored for respiration, ECG, and core temperature. A high-frequency, high-resolution digital imaging platform with linear array technology and equipped with a high-frequency linear array probe MS550D (22–55 MHz) was used throughout the study (Vevo 2100 Visual Sonics). The flow volume was modelled as a circular cylinder of length equal to the average velocity time integral and diameter measured empirically ($n = 3$ cardiac cycles), then multiplied by the heart rate (b.p.m.), then converted from $mm^3 \ min^{-1}$ to $ml \ min^{-1}$. All measurements and analyses were performed by a single blinded operator (EVK).

**Cellular studies.** Mouse inner medullary collecting duct (IMCD3) cells were obtained from ATCC. Please refer to their website for validation and mycoplasma testing (http://www.atcc.org/Products/All/CRL-2123.aspx). Cells were transfected with siRNA targeting mouse PGC1α, HCAR2 or a negative control siRNA (Qiagen) for 24 h. Niacin, mepenzolate bromide, β-hydroxybutyrate, the NAMPT inhibitor FK866 (ref. 43), and NAM were diluted to the indicated concentrations in serum-free cell culture medium. Prostaglandin E2 ($PGE_2$) was measured in the conditioned media 24–72 h after treatment.

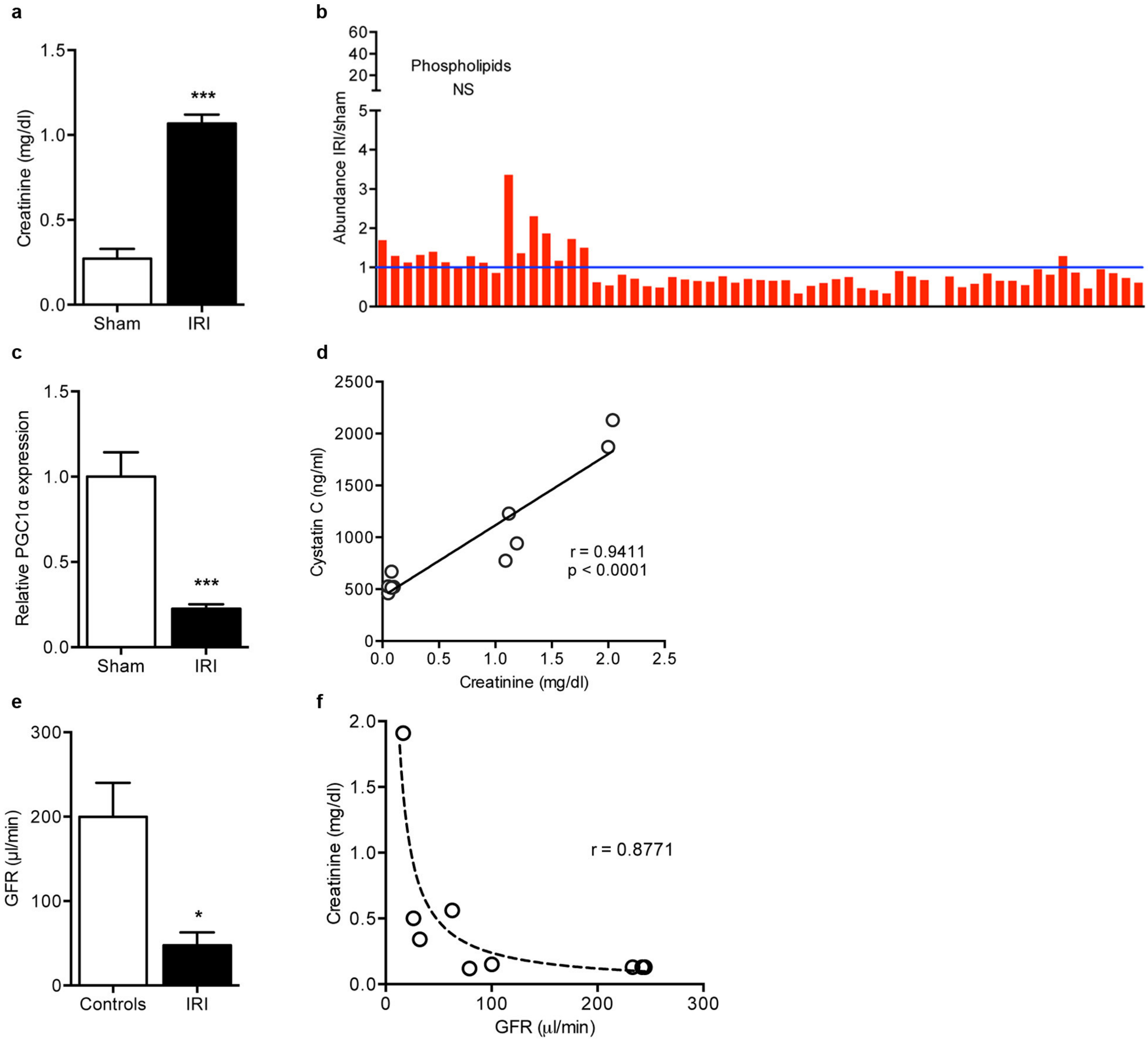
**Cystatin C.** Cystatin C in mouse serum (1:200 dilution) was measured by ELISA (R&D Systems).

**FITC-inulin clearance.** The full method is described elsewhere[44]. In brief, male C57BL/6J mice (Jackson Laboratories) were given a single bolus injection of 5%-FITC-inulin (3.74 μl per g body weight). Clearance kinetics of FITC-inulin post-injection was measured by serial blood collection at specified time points from 3 through 70 min post-injection. Blood samples were centrifuged and resulting plasma was buffered to pH 7.4 with 500 mM HEPES. Fluorescence in the buffered plasma samples was determined with 485 nm excitation, 538 nm emission. Glomerular filtration rate (GFR) was calculated from the two-phase exponential decay model outlined previously.

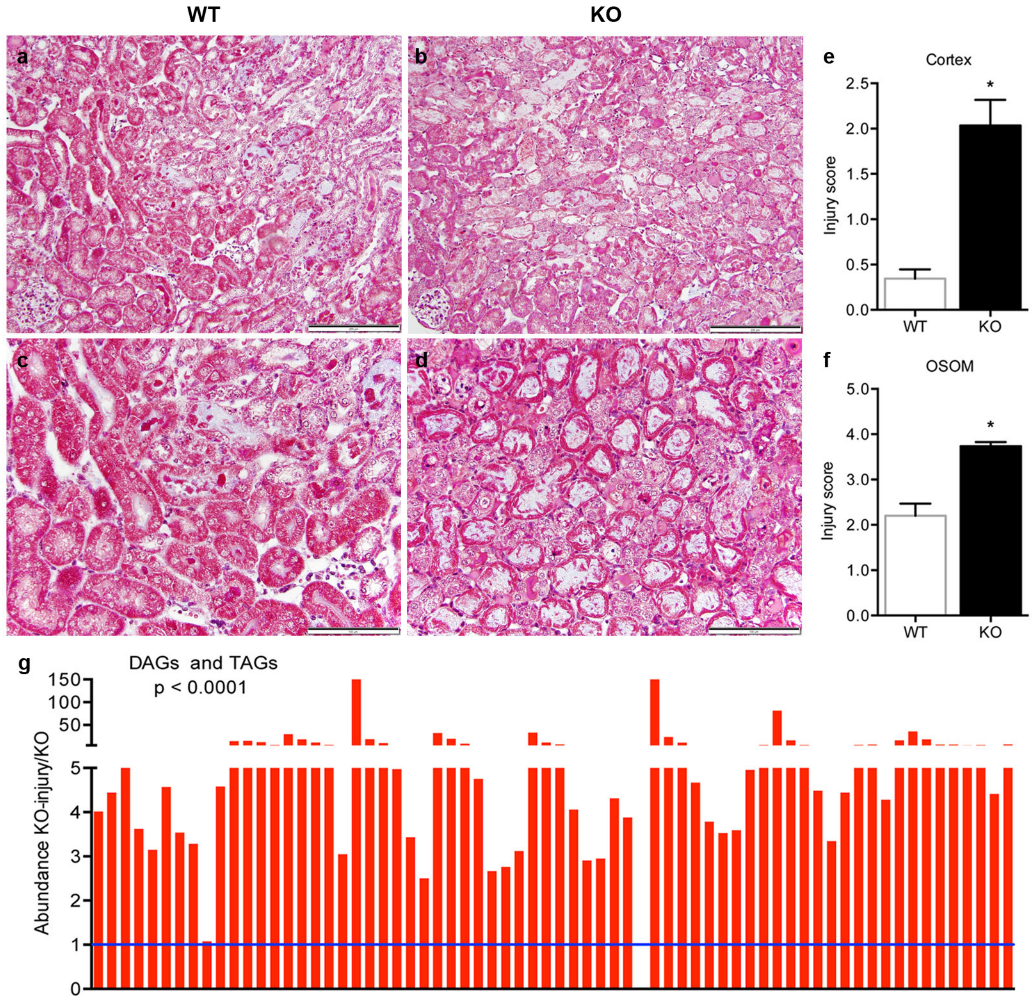
**Tissue $PGE_2$, β-OHB, and NAD measurements.** $PGE_2$ was measured in mouse kidney tissue by ELISA (Cayman Chemical). β-OHB (Cayman Chemical) and total NAD (BioVision) were measured in mouse kidney tissue by colorimetric assays. These assays were performed on kidneys used for metabolomics and lipidomics in order to compare coordinated changes in metabolism and downstream signalling. NAD measurements reflect total $NAD^+$ plus NADH.

**Statistical analysis.** Comparisons between continuous characteristics of subject groups were analysed with Mann–Whitney *U*-tests or Student's *t*-test. Survival was analysed by log-rank test. For comparisons among more than two groups, ANOVA with Bonferroni's correction was used where indicated. Associations between micro-ultrasound measurements and other functional parameters were analysed with Spearman's rank correlation coefficients. Sample size determination was guided by power calculations and prior experience. The following sample calculation was used to guide creatinine studies in mice: serum creatinine of 1.6 (±0.3 s.d.) mg $dl^{-1}$ versus 1.0 (±0.2 s.d.) requires $n = 5$ mice per condition to achieve an α-error <5% and power 96%. Mice were randomized to experimental intervention versus control. Two-tailed $P$ values < 0.05 were considered significant. Results are presented as mean ± s.e.m. and were prepared in GraphPad Prism.

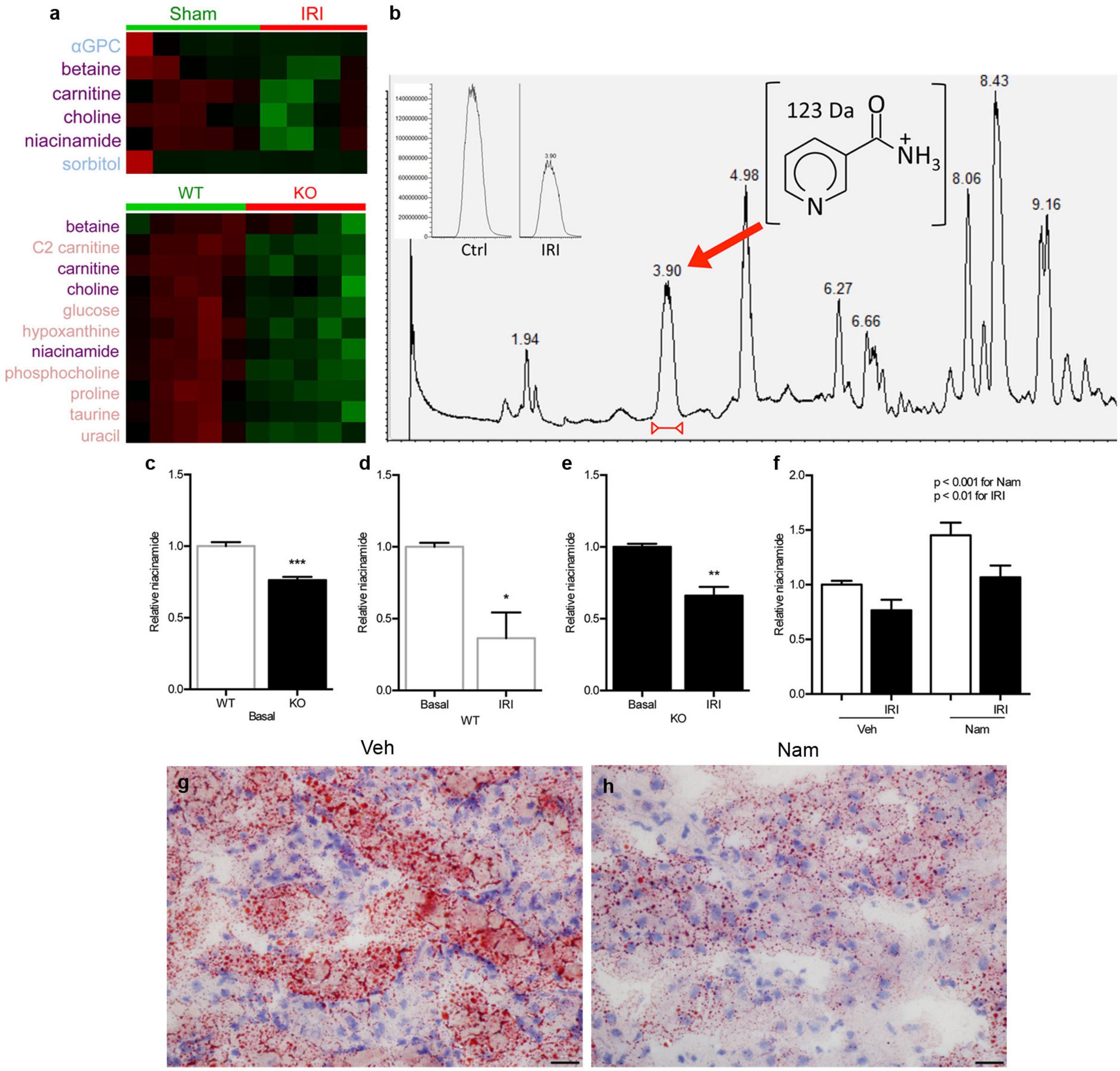
31. Tran, M. *et al.* PGC-1α promotes recovery after acute kidney injury during systemic inflammation in mice. *J. Clin. Invest.* **121,** 4003–4014 (2011).
32. Traykova-Brauch, M. *et al.* An efficient and versatile system for acute and chronic modulation of renal tubular function in transgenic mice. *Nature Med.* **14,** 979–984 (2008).
33. Rask-Andersen, M., Almen, M. S. & Schioth, H. B. Trends in the exploitation of novel drug targets. *Nature Rev. Drug Discov.* **10,** 579–590 (2011).
34. Singh, V. *et al.* Mycobacterium tuberculosis-driven targeted recalibration of macrophage lipid homeostasis promotes the foamy phenotype. *Cell Host Microbe* **12,** 669–681 (2012).
35. Feingold, K. R., Moser, A., Shigenaga, J. K. & Grunfeld, C. Inflammation stimulates niacin receptor (GPR109A/HCA2) expression in adipose tissue and macrophages. *J. Lipid Res.* **55,** 2501–2508 (2014).
36. Zsengellér, Z. K. *et al.* Cisplatin nephrotoxicity involves mitochondrial injury with impaired tubular mitochondrial enzyme activity. *J. Histochem. Cytochem.* **60,** 521–529 (2012).
37. Rhee, E. P. *et al.* A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.* **18,** 130–143 (2013).
38. Kang, C. & Ji, L. L. Muscle immobilization and remobilization downregulates PGC-1α signaling and the mitochondrial biogenesis pathway. *J. Appl. Physiol. (1985)* **115,** 1618–1625 (2013).
39. Ruas, J. L. *et al.* A PGC-1α isoform induced by resistance training regulates skeletal muscle hypertrophy. *Cell* **151,** 1319–1331 (2012).
40. Nakahata, Y., Sahar, S., Astarita, G., Kaluzova, M. & Sassone-Corsi, P. Circadian control of the NAD+ salvage pathway by CLOCK-SIRT1. *Science* **324,** 654–657 (2009).
41. Agudelo, L. Z. *et al.* Skeletal muscle PGC-1α1 modulates kynurenine metabolism and mediates resilience to stress-induced depression. *Cell* **159,** 33–45 (2014).
42. Liu, L. *et al.* Nutrient sensing by the mitochondrial transcription machinery dictates oxidative phosphorylation. *J. Clin. Invest.* **124,** 768–784 (2014).
43. Hasmann, M. & Schemainda, I. FK866, a highly specific noncompetitive inhibitor of nicotinamide phosphoribosyltransferase, represents a novel mechanism for induction of tumor cell apoptosis. *Cancer Res.* **63,** 7436–7442 (2003).
44. Qi, Z. *et al.* Serial determination of glomerular filtration rate in conscious mice using FITC-inulin clearance. *Am. J. Physiol. Renal Physiol.* **286,** F590–F596 (2004).
45. Antonica, F. *et al.* Generation of functional thyroid from embryonic stem cells. *Nature* **491,** 66–71 (2012).



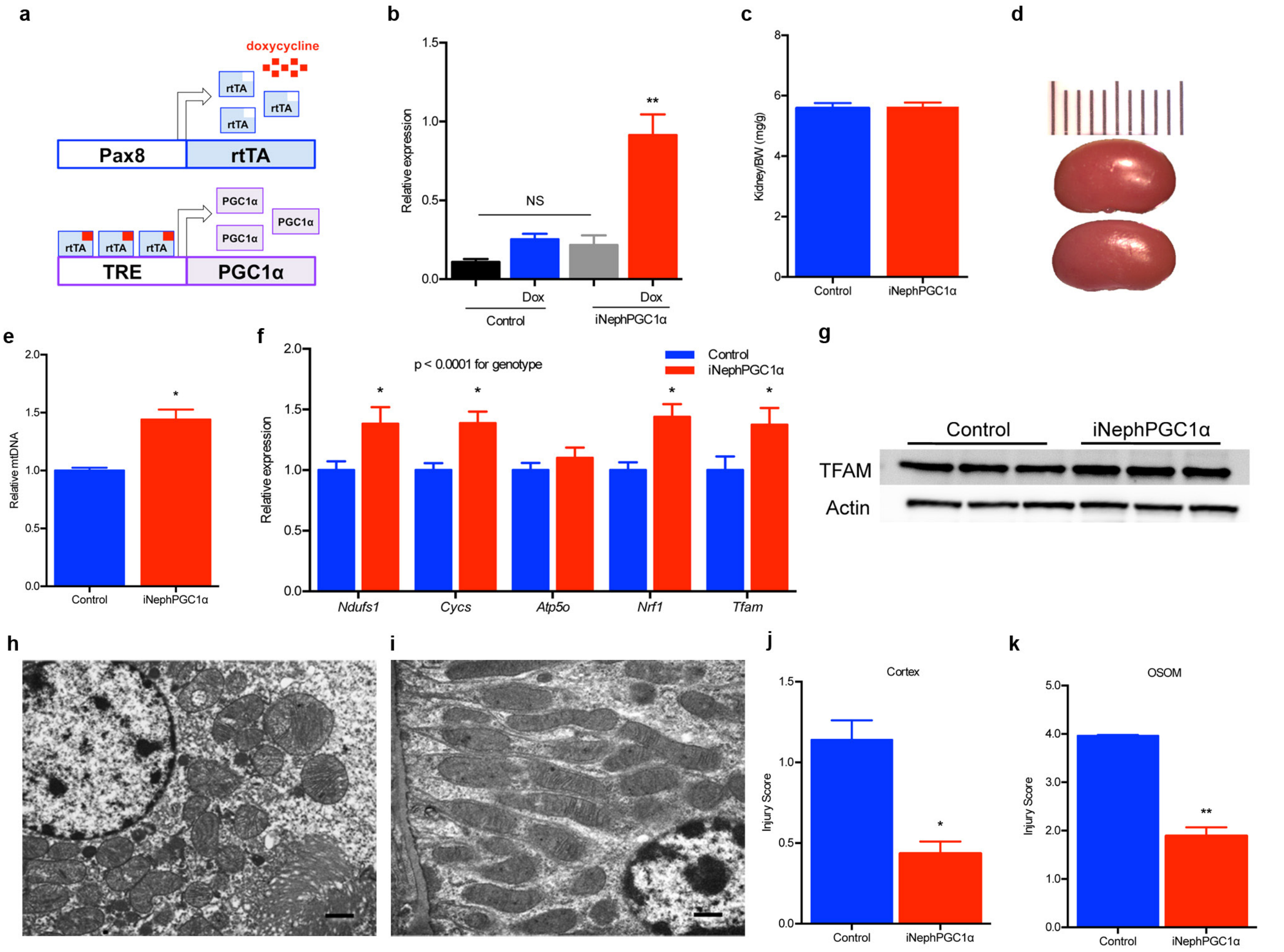
**Extended Data Figure 1 | Regulation of PGC1α and other features of post-ischaemic kidneys. a**, Serum creatinine 24 h after sham or IRI ($n = 5$ versus 14 mice); ***$P < 0.001$. **b**, Absence of class-wide changes in intrarenal phospholipids 24 h after IRI versus sham operation ($n = 6$ per group; NS, not significant). Each bar represents one lipid species. $P$ value calculated using two-way ANOVA. **c**, Renal PGC1α expression 24 h after sham or IRI ($n = 5$ animals per group); **$P < 0.01$. **d**, Correlation of LC–MS method for serum creatinine and serum cystatin C (measured by ELISA). **e**, Glomerular filtration rate in controls or 24 h after IRI was determined by two-phase exponential decay curves of fluorescently labelled inulin as described in Methods ($n = 5$ per group); *$P < 0.05$. **f**, Correlation of LC–MS method for serum creatinine with clearance of FITC–inulin. Curve fit according to formula sCr = κ/GFR where κ is a constant. Error bars, s.e.m.



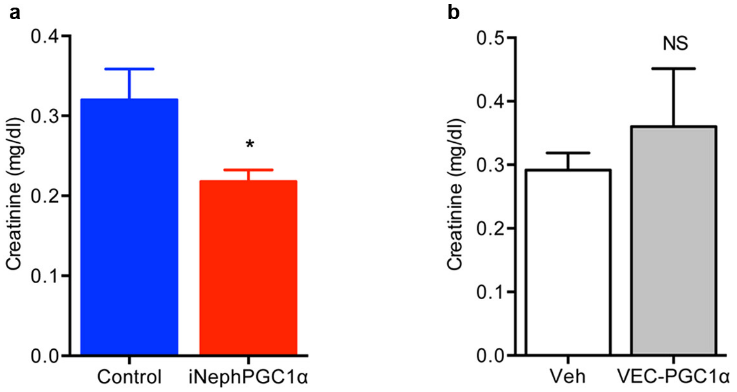
**Extended Data Figure 2 | Exacerbation of fat accumulation and tubular injury in post-ischaemic *Pgc1α*$^{-/-}$ kidneys. a–d**, Low- (top) and high- (bottom) power photomicrographs 24 h after IRI in wild-type (WT; **a**, **c**) versus *Pgc1α*$^{-/-}$ (KO; **b**, **d**) mice. Scale bars, 200 and 100 μm (top and bottom, respectively). **e**, **f**, Blinded scoring of tubular injury in cortex and outer stripe of outer medulla (OSOM) on 4-point injury scale as described in Methods ($n = 8$ wild-type versus 12 knockout mice); $^{*}P < 0.05$. **g**, Di-/tri-acylglycerols (DAGs, TAGs) in renal homogenates of knockout mice at baseline and 24 h after injury ($n = 6$ per group). Each bar represents one lipid species. *P* value calculated using two-way ANOVA. Error bars, s.e.m.



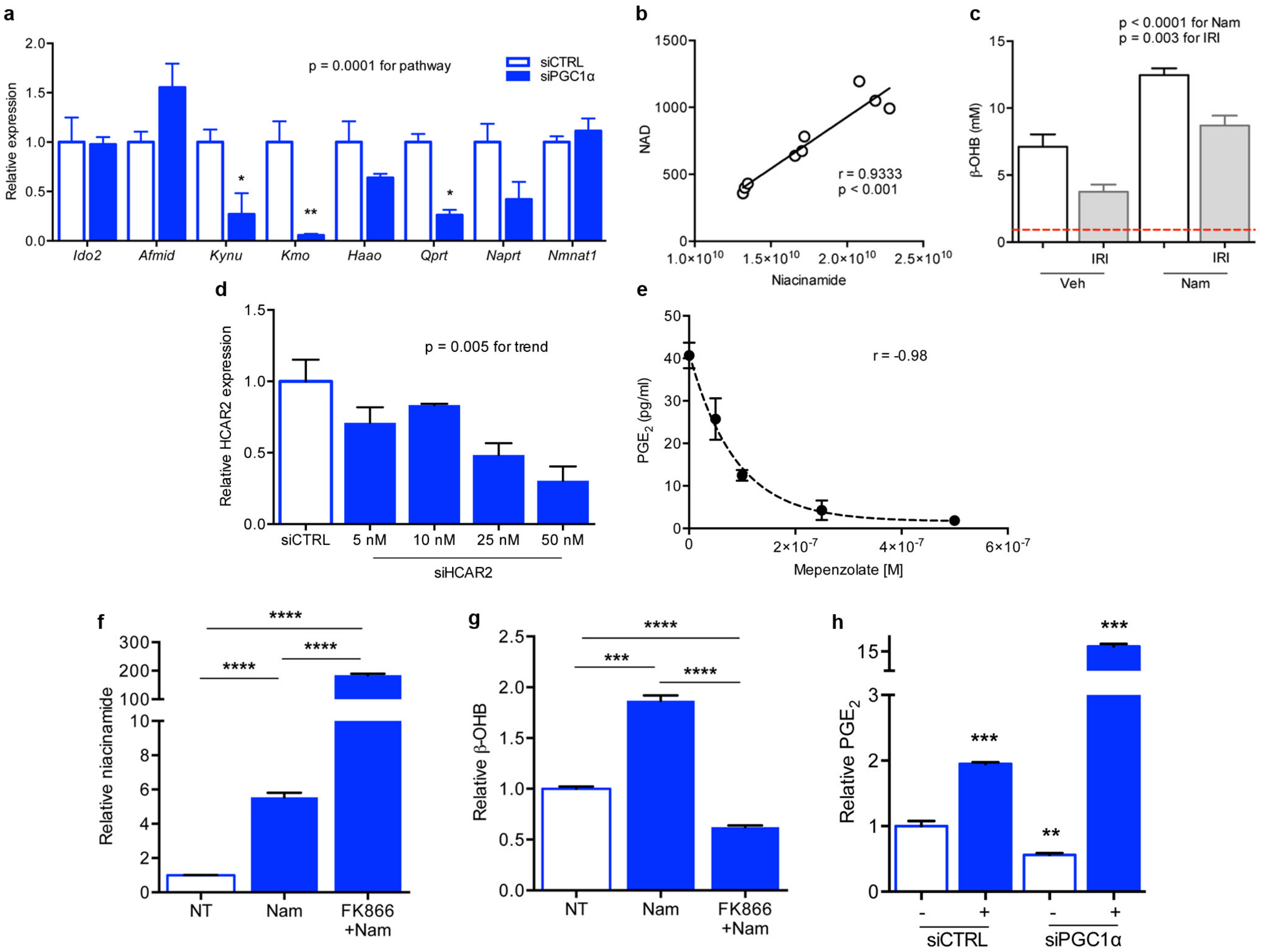
**Extended Data Figure 3 | NAM reduction from IRI and PGC1α deficiency. a**, Heat maps (red, higher; green, lower) of Bonferroni-corrected significantly different metabolites in sham versus IRI kidneys and wild-type (WT) versus knockout (KO) kidneys. Metabolites listed in purple are shared between settings. **b**, Total ion chromatogram of polar, positive-ion mode method for representative wild-type IRI sample, with NAM peak at retention time of 3.88 min. Inset shows representative NAM peaks for kidney extracts from wild-type control (Ctrl) and wild-type IRI (IRI) mice. **c**–**e**, Relative renal NAM abundance in kidneys of knockout mice versus wild-type littermates; wild-type littermates at baseline and 24 h after IRI; and knockout mice at baseline and 24 h after IRI ($n = 6$ per group). **f**, Relative renal NAM concentrations in kidneys of mice following vehicle (Veh) versus NAM treatment (400 mg kg$^{-1}$ intraperitoneal for 4 days) with and without IRI 24 h before tissue collection ($n = 6$ per group). *P* values calculated with two-way ANOVA. **g**, **h**, Oil-Red-O stain (pink) for fat accumulation 24 h after IRI with or without NAM pre-treatment (400 mg kg$^{-1}$ intraperitoneal for 4 days); scale bar, 20 μm. Error bars, s.e.m.; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$.



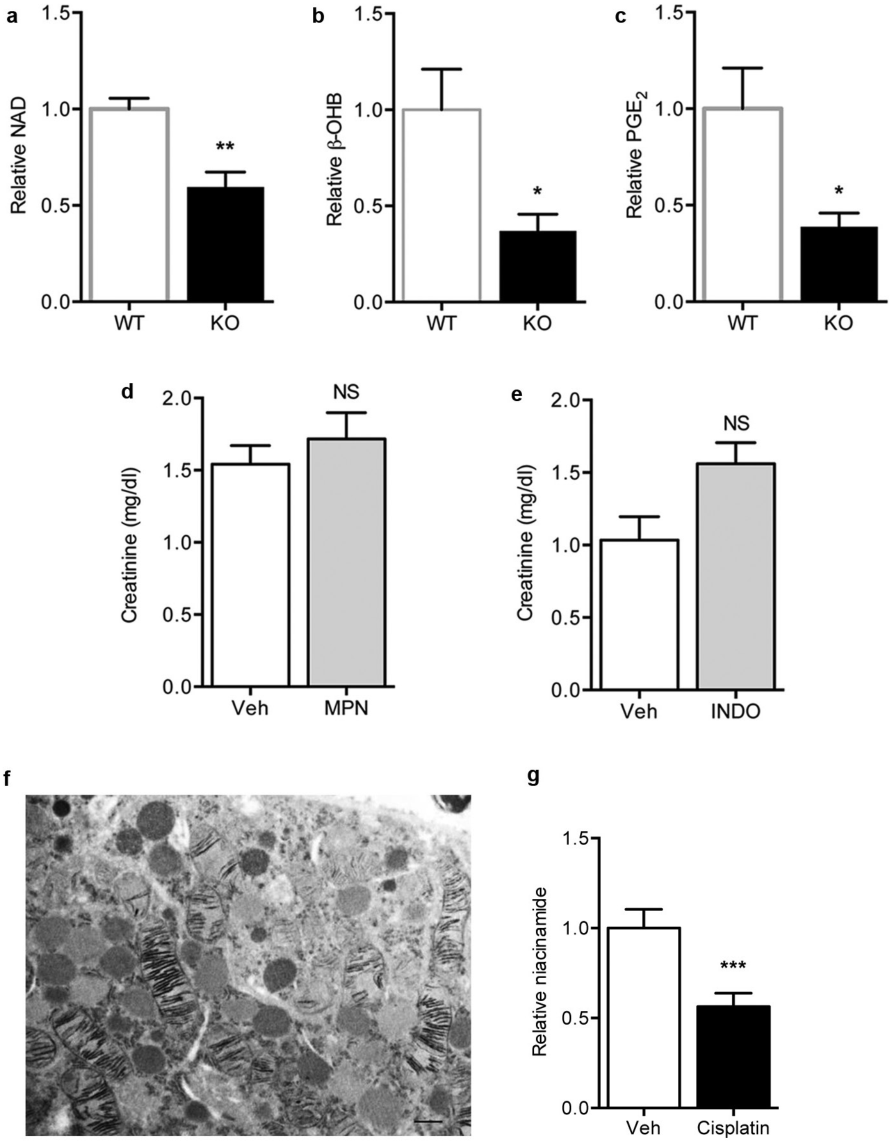
**Extended Data Figure 4 | Increased mitochondrial abundance and post-ischaemic protection in renal tubular epithelial transgenic mice (iNephPGC1α).** **a**, Schematic for generating iNephPGC1α mice. **b**, Relative renal PGC1α expression in control versus iNephPGC1α mice with and without 4 weeks of doxycycline in drinking water ($n = 5$ per group; $^{**}P < 0.01$ versus all other groups). **c**, Ratio of kidney weight to total body weight (note body weights statistically indistinguishable as well, $n = 4$ per group). **d**, Example gross images with 1 cm scale of control versus iNephPGC1α kidney. **e**, Renal mitochondrial DNA (mtDNA) copy number as described in Methods. **f**, Relative renal gene expression of PGC1α targets (*Ndufs1*, *Cycs*, *Atp5o*), partnering transcription factors (*Nrf1*), and the mitochondrial transcription factor, *Tfam*. Results analysed by two-way ANOVA with *P* value for genotype as noted. $n = 8$ per group. $^{*}P < 0.05$ versus control after Bonferroni correction. **g**, Western blot analysis of kidney lysates for transcription factor a, mitochondrial (TFAM)[38] and loading control. **h**, **i**, Transmission electron microscopy of mitochondria sectioned perpendicular and parallel to long axis demonstrating normal morphology in iNephPGC1α mice (representative of $n = 4$ per group); scale bar, 500 nm. **j**, **k**, Blinded scoring of tubular injury in cortex and outer stripe of outer medulla ($n = 8$ control; 12 iNephPGC1α). Error bars s.e.m.; $^{*}P < 0.05$, $^{**}P < 0.01$; NS, not significant.



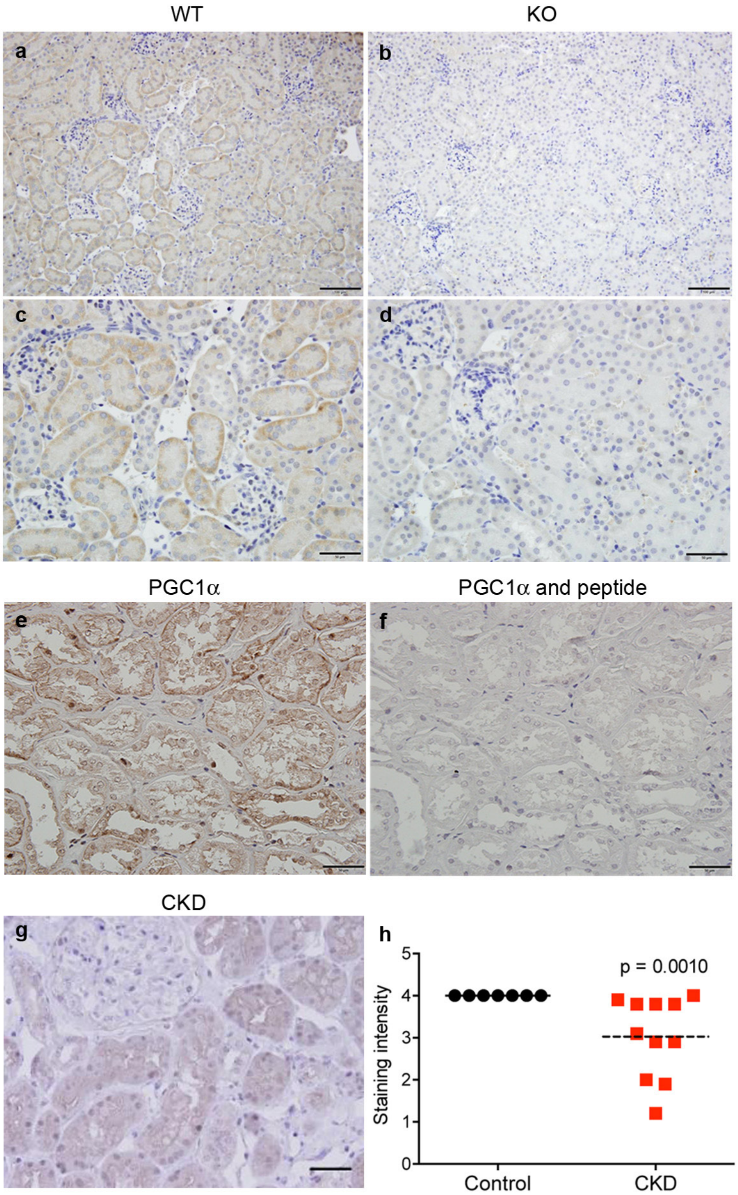
**Extended Data Figure 5 | Renal protection in systemic inflammation conferred by renal tubular epithelial, but not endothelial, PGC1α.** **a**, Serum creatinine 24 h after bacterial endotoxin injection (LPS O111:B4), $n = 9$ per group. **b**, Serum creatinine 24 h after bacterial endotoxin (LPS O111:B4) in endothelial-specific (VEC, VE-cadherin) PGC1α transgenic mice (VEC-tTA × TRE-PGC1α), $n = 5$ per group. Error bars, s.e.m., $*P < 0.05$; NS, not significant.



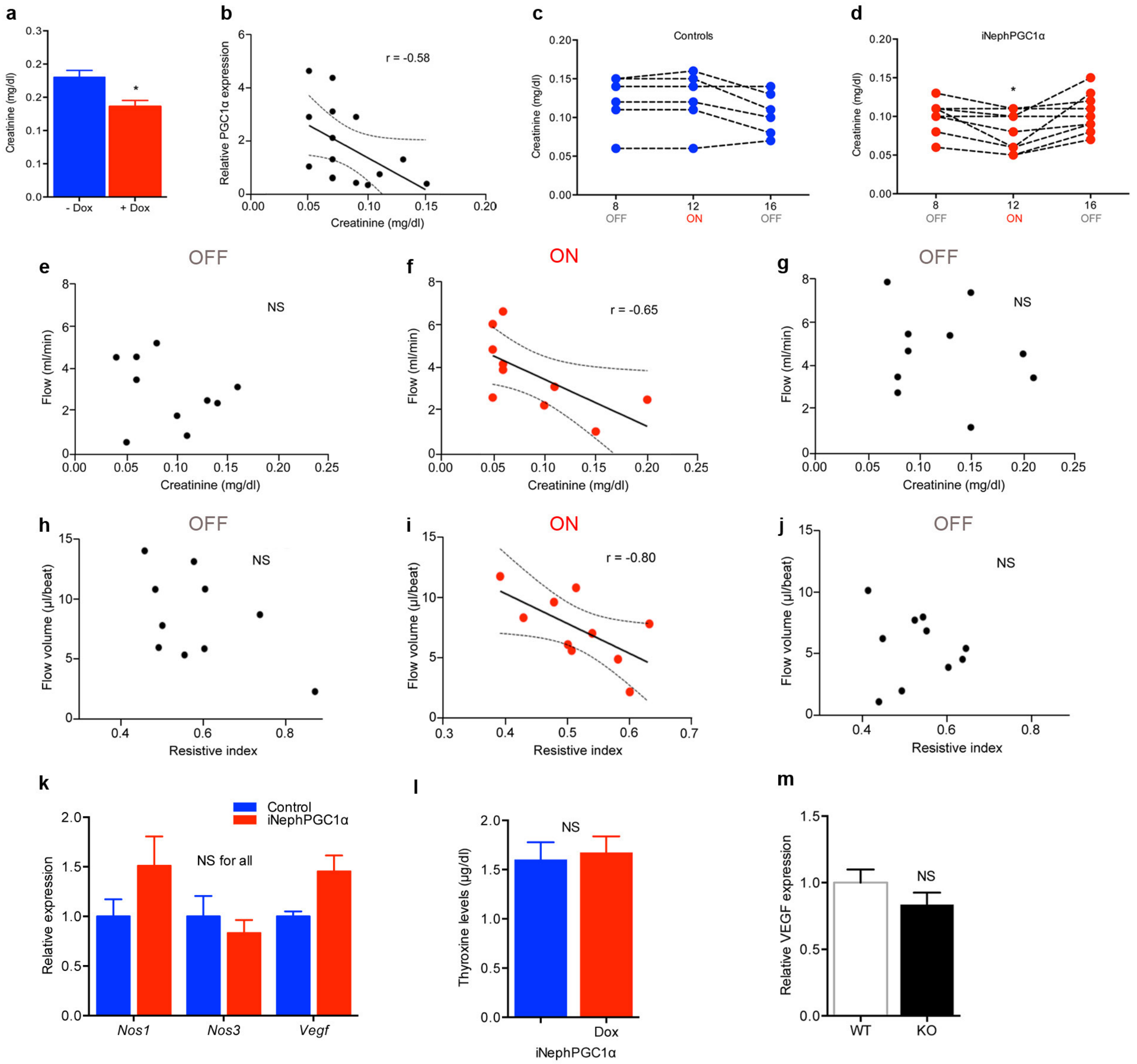
**Extended Data Figure 6 | PGC1α-dependent *de novo* NAD biosynthesis and NAD-dependent accumulation of β-OHB and $PGE_2$.** **a**, Gene expression for *de novo* NAD biosynthetic pathway in control renal tubular cells versus 48 h after PGC1α knockdown ($n = 3$ per condition). The gene expression set corresponds to the eight transcripts whose abundance was measured in kidney homogenates in Fig. 3. $P = 0.0001$ by two-way ANOVA with Bonferroni-corrected comparisons as indicated. **b**, Correlation of renal NAM versus renal NAD in mice treated with vehicle or different doses of NAM (one intraperitoneal dose of 100–400 mg $kg^{-1}$). Arbitrary units on *x* and *y* axes. **c**, Renal β-OHB concentrations in kidneys of mice following vehicle (Veh) versus NAM treatment (400 mg $kg^{-1}$ intraperitoneal for 4 days) with and without IRI 24 h before tissue collection ($n = 5$ per group). *P* value calculated with two-way ANOVA. Dashed line indicates normal circulating concentration of β-OHB. **d**, Dosing for siRNA against HCAR2 in renal tubular cells. **e**, Dose–inhibition curve in renal tubular cells for $PGE_2$ release following 24 h of mepenzolate bromide at the indicated concentrations ($n = 3$ replicates per concentration)[33–35]. **f**, **g**, Intracellular NAM and secreted β-OHB for renal tubular cells following treatment with NAM (1 μM for 24 h) with or without pre-treatment with the NAMPT inhibitor FK866 (10 nM, $n = 6$ per condition). **h**, $PGE_2$ in conditioned media of renal tubular cells after control versus PGC1α knockdown and with and without exogenous β-OHB application (+, 5 mM, $n = 6$ per condition, *P* values versus control group). Error bars, s.e.m.; $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$.



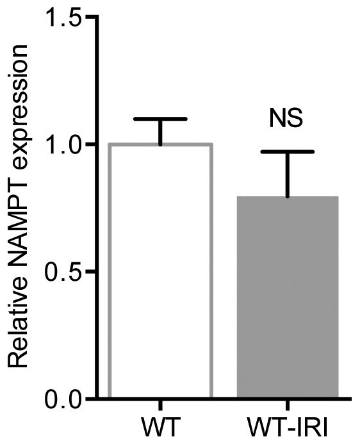
**Extended Data Figure 7 | Effects of PGC1α on renal metabolites and features of cisplatin nephrotoxicity. a–c**, Relative renal NAD, β-OHB, and PGE$_2$ concentrations in wild-type (WT) littermates versus *Pgc1α*$^{-/-}$ (KO) mice (*n* = 6 per group). **d**, Serum creatinine in genetic control mice for iNephPGC1α 24 h after IRI with vehicle versus mepenzolate (MPN, 10 mg kg$^{-1}$ intraperitoneal) treatment (*n* = 5 per group). **e**, Serum creatinine in genetic control mice for iNephPGC1α 24 h after IRI with vehicle versus indomethacin (INDO, 10 mg kg$^{-1}$ intraperitoneal) treatment (*n* = 6 per group). **f**, Transmission electron microscopy with cytochrome *c* oxidase enzyme histochemistry of proximal tubular cell 24 h following cisplatin exposure (25 mg kg$^{-1}$ intraperitoneal) demonstrating mitochondrial injury. Scale bar, 500 nm. **g**, Relative renal NAM concentrations following cisplatin as in **f**. Error bars, s.e.m.; *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$; NS, not significant.



**Extended Data Figure 8 | Renal immunostaining for PGC1α declines in human chronic kidney disease. a–d**, Low- (**a**, **b**) and high- (**c**, **d**) power photomicrographs of PGC1α immunoreactivity (brown) in wild-type (WT, left) littermates and *Pgc1α*$^{-/-}$ (KO, right) kidneys. Scale bars, 100 and 50 μm. **e**, **f**, Representative results of peptide competition attenuating PGC1α immunoreactivity against human kidney ($n = 4$) as described in Methods. **g**, Representative immunostaining (brown) for PGC1α in a renal biopsy with chronic kidney disease (CKD). Scale bar, 50 μm. **h**, Results of scoring PGC1α immunostaining intensity (1, weakest; 4, strongest) in specimens with CKD by blinded operator. Each dot represents a unique specimen. Analysed with Mann–Whitney *U*-test.



**Extended Data Figure 9 | Evidence for renal-tubular-epithelial-PGC1α-dependent reversible vascular relaxation.** **a**, Serum creatinine in uninduced (−Dox) versus induced (+Dox) iNephPGC1α mice ($n = 8$ mice per group). **b**, Comparison of serum creatinine with degree of renal PGC1α expression, $P < 0.05$. **c**, **d**, Serial serum creatinines in iNephPGC1α mice versus controls before PGC1α induction (OFF ), after 4 weeks of PGC1α induction (ON), and after 4 weeks of washout (OFF), $n = 5$ per group; $^*P < 0.05$ as calculated using repeated-measures ANOVA. **e**–**g**, Comparison of serum creatinine at different time points with renal artery flow in iNephPGC1α mice from **d**, $P < 0.05$ when correlation coefficient $r = -0.65$. **h**–**j**, Comparison of resistive index with renal artery flow volume in iNephPGC1α mice from **d**, $P < 0.05$ when correlation coefficient $r = -0.80$. **k**, Relative renal expression of VEGF and nitric oxide synthases 1 and 3 ($n = 6$ per group). Analysed by two-way ANOVA with Bonferroni corrections. **l**, Circulating thyroxine levels in iNephPGC1α mice with and without gene induction ($n = 5$ per group) to rule out Pax8-related thyrotoxicosis driving perfusion differences as previously described[45]. **m**, Relative renal expression for VEGF in *Pgc1α*$^{-/-}$ mice (KO) versus wild-type (WT) littermates ($n = 6$ per group). Error bars, s.e.m.; NS, not significant.



**Extended Data Figure 10 | Relative renal expression for NAMPT in wild-type (WT) mice before and 24 h after IRI ($n = 6$ per group).** Error bars, s.e.m.; NS, not significant.

doi:10.1038/nature16993

# Mycocerosic acid synthase exemplifies the architecture of reducing polyketide synthases

Dominik A. Herbst[1]*, Roman P. Jakob[1]*, Franziska Zähringer[1]† & Timm Maier[1]

**Polyketide synthases (PKSs) are biosynthetic factories that produce natural products with important biological and pharmacological activities[1–3]. Their exceptional product diversity is encoded in a modular architecture. Modular PKSs (modPKSs) catalyse reactions colinear to the order of modules in an assembly line[3], whereas iterative PKSs (iPKSs) use a single module iteratively as exemplified by fungal iPKSs (fiPKSs)[3]. However, in some cases non-colinear iterative action is also observed for modPKSs modules and is controlled by the assembly line environment[4,5]. PKSs feature a structural and functional separation into a condensing and a modifying region as observed for fatty acid synthases[6]. Despite the outstanding relevance of PKSs, the detailed organization of PKSs with complete fully reducing modifying regions remains elusive. Here we report a hybrid crystal structure of *Mycobacterium smegmatis* mycocerosic acid synthase based on structures of its condensing and modifying regions. Mycocerosic acid synthase is a fully reducing iPKS, closely related to modPKSs, and the prototype of mycobacterial mycocerosic acid synthase-like[7,8] PKSs. It is involved in the biosynthesis of $C_{20}$–$C_{28}$ branched-chain fatty acids, which are important virulence factors of mycobacteria[9]. Our structural data reveal a dimeric linker-based organization of the modifying region and visualize dynamics and conformational coupling in PKSs. On the basis of comparative small-angle X-ray scattering, the observed modifying region architecture may be common also in modPKSs. The linker-based organization provides a rationale for the characteristic variability of PKS modules as a main contributor to product diversity. The comprehensive architectural model enables functional dissection and re-engineering of PKSs.**

Each homodimeric PKS module elongates acyl-carrier protein (ACP) tethered precursors by the sequential action of an acyltransferase (AT) and a ketosynthase (KS), organized in the essential condensing region (KS–AT). The product can further be sequentially modified by a ketoreductase (ΨKR/KR), a dehydratase (DH), and an enoylreductase (ER)[1,3]. These optional domains form the variable modifying region of PKSs. Mycocerosic acid synthase (MAS) is a fully reducing PKS with a complete modifying region (DH–ΨKR–ER–KR). It iteratively elongates linear $C_{12}$–$C_{20}$ starter fatty acids in one to four rounds with methyl-malonyl-CoA extender units[8] to produce mycocerosic acids. These MAS products form the core of phenolic glycolipids and phthiocerol dimycocerosates, key lipids of the mycobacterial cell envelope[8]. The condensing and modifying regions of MAS are centrally connected by non-conserved linkers, which permit large-scale relative motions in related systems[10]. To obtain a high-quality hybrid model, we divided MAS into its condensing and modifying regions, and excluded the flexibly tethered ACP (Fig. 1a).

Three constructs of staggered carboxy (C)-terminal length were employed to define the length of the condensing region (see Methods). All variants crystallized under the same condition; structure determination mapped the last ordered residue to Glu887. The structure of the most extended variant (1–892) was refined at 2.2 Å resolution (Extended Data Table 1a). MAS KS–AT comprises an α/β-fold linker domain (LD) connecting AT to KS (Fig. 1b). The monomeric condensing region closely resembles those of other PKSs and fatty acid synthases (FASs)[6,11,12], the closest structural homologue at individual domain level is module 5 of the 6-deoxyerythronolide B synthase (DEBS) PKS (Extended Data Table 2a). Compared with previous KS–AT di-domain structures, the AT domain is slightly rotated towards the C-terminal post-AT linker.

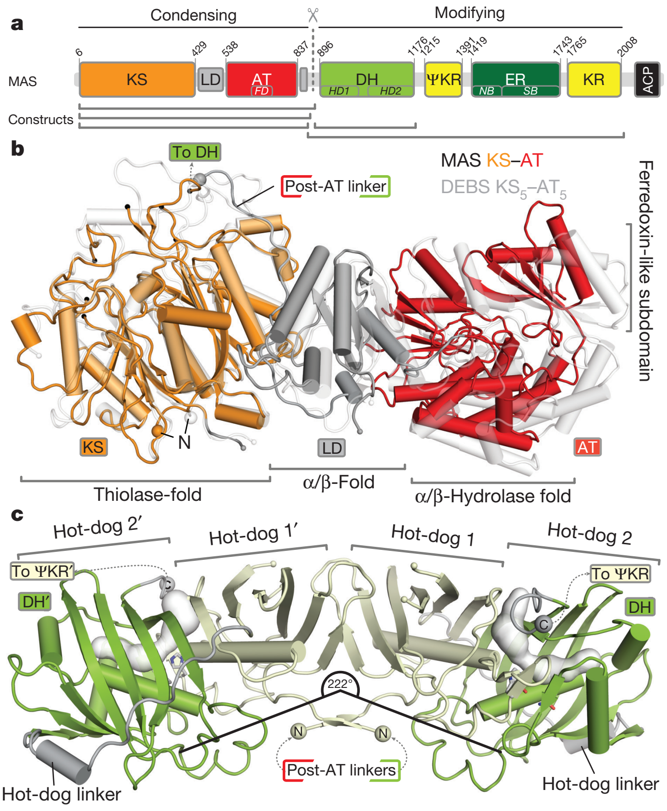


**Figure 1 | Domain organization, condensing region, and dimeric DH domain of MAS. a**, MAS is organized in a condensing KS–AT ($AT_{FD}$, ferredoxin-like AT subdomain) and a modifying DH–ΨKR–ER–KR region ($DH_{HD1/2}$, hot-dog fold 1 or fold 2 of DH; ΨKR, non-catalytic pseudo-KR domain; $ER_{NB/SB}$, nucleotide/substrate binding ER subdomain), followed by a flexibly tethered ACP domain. Crystallized constructs are indicated. **b**, Monomeric condensing region crystal structure. The AT position corresponds to a rotation around a hinge in the LD relative to DEBS $KS_5$–$AT_5$ (ref. 11) (white). Black spheres indicate ends of disordered segments (aa 47–65, 132–151, 211–220, 277–283). **c**, Crystal structure of the dimeric DH. Each monomer comprises two hot-dog folds connected by a 20-aa hot-dog linker (grey). The DH active-site tunnel (white) has two openings and the dimer is bent with an interdomain angle of 222°.

[1]Department Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland. †Present address: F. Hoffmann-La Roche AG, Grenzacherstrasse 124, 4070 Basel, Switzerland.
*These authors contributed equally to this work.

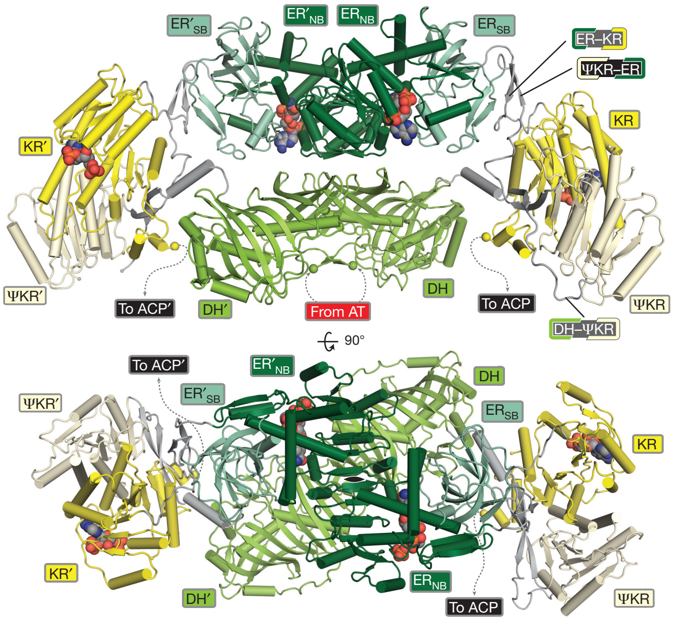


**Figure 2 | Crystal structure of the dimeric MAS modifying region.** The MAS modifying region is organized by the dimerization of the central DH (light green) and ER (darker greens) domains (upper panel: front view, lower panel: top view). The DH dimer reveals virtually the same bent organization as observed in the crystal structures of the isolated DH domains. The ΨKR/KR (yellow) domains are laterally tethered, share no direct interface with any other catalytic domain, and their positioning is the most variable of all domains. Bound cofactors are shown in sphere representation coloured by element type. A two-fold dimer axis is indicated in the lower panel.

Isolated MAS KS–AT crystallized as a monomer lacking the canonical KS-based dimerization[6,11,12], but is in monomer–dimer equilibrium in solution with a dissociation constant, $K_d$, of 0.4 mM as determined by analytical ultracentrifugation. It is the first condensing region crystallized as monomer, but dissociation has been observed for other condensing region fragments in the absence of dimeric partner domains[12–14]. Differences to canonical dimeric KS, as exemplified by DEBS $KS_5$ (ref. 11) or CurL KS[12], are observed around the dimer interface, presumably because of the absence of stabilizing dimer interactions: the interface-spanning active-site tunnel is incomplete and the loop containing the catalytic cysteine (Cys178) is bent outwards by 9 Å into a non-productive conformation, while the active-site histidines (His313, His349) are at expected positions (Extended Data Fig. 1a–c). Four interface segments of 6–19 amino-acid (aa) length are disordered (Fig. 1b), while equivalent regions are ordered in dimeric KS domains.

A single mode of dimerization based on canonical KS organization was identified by automated sequence-based methods (see Methods), and homology-based modelling of dimeric MAS KS–AT restores the active-site tunnel and a productive conformation of Cys178 (Extended Data Fig. 1b, c). The KS–AT dimer adopts a linear shape owing to the rotation of AT relative to KS (Extended Data Fig. 1d). The C-terminal post-AT linkers of the condensing region, which connect to the modifying region, are proximal to the two-fold dimer axis above the KS active site, as observed in previous condensing region structures[6,11,12].

The DHs connect the modifying region to the post-AT linkers of the condensing region. We solved crystal structures of a MAS DH construct (aa 884–1186) (Extended Data Table 1a), which overlaps in sequence with the crystallized KS–AT, in two crystal forms with a total of six protomers arranged into almost identical dimers. The DH protomer is composed of two hot-dog folds connected by a 20-aa hot-dog linker (Fig. 1c). A hydrophobic substrate binding tunnel extends over both hot-dog folds with entrances near the C terminus and at the distal end of hot-dog fold 2. Active-site residues are contributed by both hot-dog folds and are located close to the C terminus (Extended Data Fig. 1e). The nearest structural homologues of DH protomers are modPKS DH domains (Extended Data Table 2a). In the DH dimer, the two protomers arrange with their lateral ends bent towards the post-AT linkers with an interdomain angle of 222° (Fig. 1c). The MAS DH dimer is distinct from the V-shaped DH arrangement in FAS[6], which lacks a dimerization interface and is bent into the opposite direction at an angle of 96°. MAS DH rather resembles linear DH dimers of modPKSs with interdomain angles of 167–203° (refs 15–17) and a common mode of dimerization via 'handshake' interactions between β-strands of the amino (N)-terminal hot-dog folds (Extended Data Fig. 1f–h).

To obtain an authentic representation of the MAS modifying region, we crystallized the complete DH–ΨKR–ER–KR segment in presence of $NADP^+$, which is dimeric in solution based on analytical ultracentrifugation. On the basis of small-angle X-ray scattering (SAXS), ACP deletion does not affect the overall structure of this region (Extended Data Fig. 2a–c). The crystallographic asymmetric unit reveals a complex packing of nine dimers related by non-crystallographic symmetry (NCS). The corresponding 18 polypeptide chains comprise 20,502 aa (2.2 MDa protein mass), of which 17,680 are modelled. Real-space NCS averaging and NCS-restrained refinement led to a high-quality model ($R_{work}/R_{free} = 0.23/0.24$) at 3.75 Å resolution (Fig. 2, Extended Data Table 1a and Extended Data Fig. 2d–f). The modifying region dimerizes along an extended interface formed by DH and ER (Extended Data Table 2b); the ΨKR/KR is laterally connected to DH and ER. MAS, as well as most reducing modPKSs, lacks a non-catalytic pseudo-methyltransferase domain (ΨME), which is a characteristic of FASs and fiPKSs. The DH in the modifying region adopts the same dimeric structure as in the isolated form (Extended Data Table 2a), demonstrating the intrinsic nature of DH dimerization and its role in organizing the modifying region. The ER domain is characterized by a large active-site tunnel and a well-ordered $NADP^+$ cofactor (Extended Data Fig. 3a, b). The ER dimerizes via pseudo-continuous β-sheet formation between the nucleotide binding subdomains ($ER_{NB}$) and provides the largest contribution to the modifying region dimer interface. Its closest structural neighbours are the isolated modPKS ERs from *Lyngbya majuscula*[18] and the SpnB ER–ΨKR/KR di-domain[19] (Extended Data Table 2a), even though these ERs are monomeric. The dimerization mode of MAS ER closely resembles those of the $ER_{NB}$ subdomain of the PpsC modPKS (Protein Data Bank accession number 1PQW) and the ER of FAS[6,20] (Extended Data Fig. 3c). The split ΨKR/KR resembles modPKSs ΨKR/KR (Extended Data Table 2a)[21]; as in related B-type KR domains[22], a flexible lid region (aa 1948–1960) remains disordered in the absence of ligand, and concomitantly, the nicotinamide moiety of $NADP^+$ is disordered (Extended Data Fig. 3d). The MAS ΨKR exhibits an N-terminal β–α–β–α extension, which is commonly observed in modPKSs, but not in FASs[6,23]; this extension exhibits increased flexibility, as indicated by temperature factor distributions (Extended Data Fig. 3e, f).

Previously, modifying region architecture was discussed on the basis of domain interfaces in FAS and PKSs fragments[1]. However, the current analysis of the MAS modifying region reveals a striking absence of stable interfaces between the different domains: the ER dimer rests on a platform formed by the DH dimer, but the interface between the two is small and variable (345–638 $Å^2$) (Extended Data Table 2b and Supplementary Video 1). The ΨKR/KR does not contact its neighbouring domains at all and is the region of highest structural variability. Instead, the architecture of the modifying region is based on three linkers interconnecting the ΨKR/KR, DH, and ER domains, which act as spacers as well as interaction partners among each other and with catalytic domains (Fig. 3a, b and Extended Data Fig. 2d), as follows. (1) The 27-aa ΨKR–ER linker plays a central organizing role by forming extended interfaces to ΨKR/KR (975 ± 28 $Å^2$) and ER

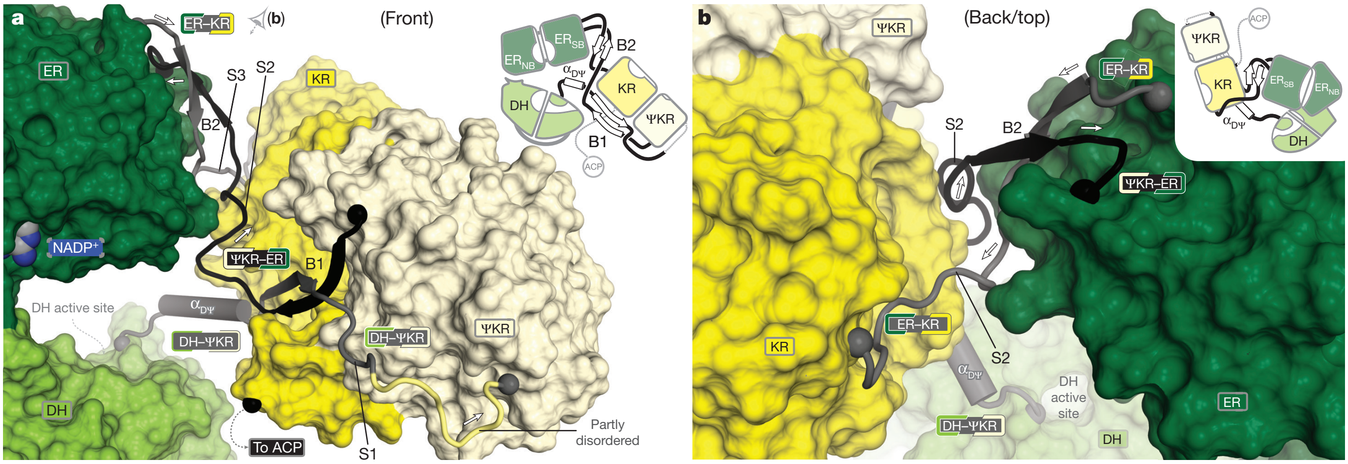


**Figure 3 | Linker-based organization of the MAS modifying region.** The DH and ER lack direct interdomain contacts to the ΨKR/KR domains. The modifying region is established by an interplay of irregular and helical linker segments with two double-stranded antiparallel linker β-sheets (B1, B2), which interact with the ΨKR/KR and ER domain, respectively. **a**, The DH–ΨKR linker (aa 1177–1214) provides helix $\alpha_{D\Psi}$ as a spacer between DH and β-sheet B1 on the surface of the ΨKR/KR. The DH–ΨKR linker continues into segment S1 and ends in a partly disordered loop (light yellow), which was traced only in one chain. The central ΨKR–ER linker (aa 1392–1418) engages in both β-sheets (B1 and B2); the stretch S2 between sheet B1 and B2 adopts two alternative conformations among different chains. **b**, The ER–KR linker (aa 1744–1764) contains an irregular stretch (S3, aa 1753–1764), which is considerably longer than required to bridge the interdomain gap.

(353 ± 20 $\text{Å}^2$). Moreover, it interacts with each of the other two linkers via two double-stranded antiparallel β-sheets. The β-sheet formed between the ΨKR–ER and DH–ΨKR linkers (B1 in Fig. 3) is embedded in a surface groove of the ΨKR/KR. It partly extends the Rossman-fold of the KR and is conserved in PKSs (Extended Data Fig. 4). The β-sheet between the ER–KR and ΨKR–ER linker (B2 in Fig. 3) mostly interacts with the ER and establishes a gap between the ER and KR. (2) The 38-aa DH–ΨKR linker comprises an N-terminal 10-aa α-helix ($\alpha_{D\Psi}$ in Fig. 3) followed by the β-strand paired to ΨKR–ER linker and an irregular segment (S1 in Fig. 3a), which wraps around the ΨKR. Helix $\alpha_{D\Psi}$ separates DH and ΨKR/KR; fragments of it are also observed in structures of isolated DH domains from the Curacin PKS (CurH, K, J)[15]. (3) The 20-aa ER–KR linker consists of a terminal irregular segment (S3 in Fig. 3) and the central β-strand paired to the ΨKR–ER linker. It contacts ER and KR via interfaces of 432 ± 24 $\text{Å}^2$ and 547 ± 14 $\text{Å}^2$, respectively, and together with the ΨKR–ER linker forms a continuous connection layer between these domains.

To obtain a MAS hybrid model we connected the overlapping modifying and condensing region fragments *in silico* (Fig. 4a). We assume that the condensing region adopts a canonical dimeric state upon tethering to the dimeric modifying domain. The relative orientation of the condensing and modifying regions is not defined by the two structures and was chosen in accordance to intact FAS[6]. As in FAS, the two fragments connect without secondary contacts outside the linking region. On the basis of multiple modes of motion around the central linkage observed in FAS[10], the selected orientation may represent only one out of an ensemble of states in both multienzymes. Helix formation of the sequence segment linking modifying and condensing domain was observed at the N terminus of four protomers in the crystallized modifying region under stabilization by crystal contacts. The central connection in the hybrid model consequently was modelled with an α-helix (Extended Data Fig. 5a, b), in contrast to an irregular linker in FAS. Notably, short helices in equivalent sequence positions are observed in modPKS DEBS $DH_4$ (ref. 16) and Rif $DH_{10}$ (ref. 17) as well as in RhiE KS-B[24] (Extended Data Fig. 5c, d), suggesting a more general conservation of helical linkers in modPKSs.

Conformational dynamics are a key component of multi-enzyme action. They have been visualized by EM for FAS[10] and PikAIII[13,25], but not at resolutions required for mechanistic dissection. The crystallographic visualization of 18 instances of the modifying regions now provides an opportunity to analyse conformational variability in MAS. The central DH and ER dimers each behave as rigid bodies, but the ERs move in a screw motion with a translation of up to 8.5 Å and a rotation by 13.6° on the DH platform (Fig. 4b and Extended Data Fig. 6a, b). The ERs are conformationally coupled to the ΨKR/KRs (Supplementary Video 1). Owing to the tethering of ΨKR/KR to both the DH and ER, the screw motion of the ERs is transduced into a rotation of the ΨKR/KRs by up to 40° via a pivot in the linkers (Extended Data Fig. 6c, d). Even larger motions may occur in solution, as indicated by pronounced disorder of some ΨKR/KRs in the crystal. Importantly, conformational coupling via relative DH–ER motions provides crosstalk between the two lateral clefts of MAS. Although a mechanism for reading out active-site states remains unknown, this coupling could transmit reaction states across the MAS dimer. Notably, the mobile ACP is tethered to the most flexible catalytic domain (ΨKR/KR), creating a hierarchical network of gradually increasing domain flexibility.

Only one condensing region instance has been visualized here, but it extends the previously observed range of KS–AT conformations[6,11,12] (Extended Data Fig. 6e, f). MAS KS–AT features the most linear conformation, which results in narrowing the gap to the modifying region and shortening of the AT–ACP anchor distances. Variations between condensing regions correspond to a hinge-bending motion of AT around a pivot in LD (Supplementary Video 2 and Fig. 4b). Although experimental evidence of flexibility in each system is lacking, normal-mode analysis indicates a conservation of this hinge in all KS–AT di-domains. In the EM reconstruction of PikAIII the AT domain is rotated by approximately 90° relative to MAS and remains a clear outlier to the set of KS–AT regions depicted by crystallography[6,11,12], EM[10], and SAXS[14].

The MAS hybrid model is a prototype for mycobacterial MAS-like (Msl-)PKS organization[7]. Moreover, our structural data reinforce the sequence-based conclusion that MAS also serves as a model for modPKSs. Despite its iterative mode of action, MAS is clearly assigned phylogenetically to modPKSs (27–35% sequence identity) rather than fiPKSs (20–22% identity) or FASs (19% identity) (Extended Data Fig. 7). Structurally, the closest neighbours of all individual MAS domains are from modPKSs. The absence of a ΨME domain and the presence of a ΨKR β–α–β–α extension distinguish MAS and most modPKSs from FASs and fiPKSs. 'Handshake' interactions of isolated dimeric DHs are observed only in modPKSs, but not in FASs.

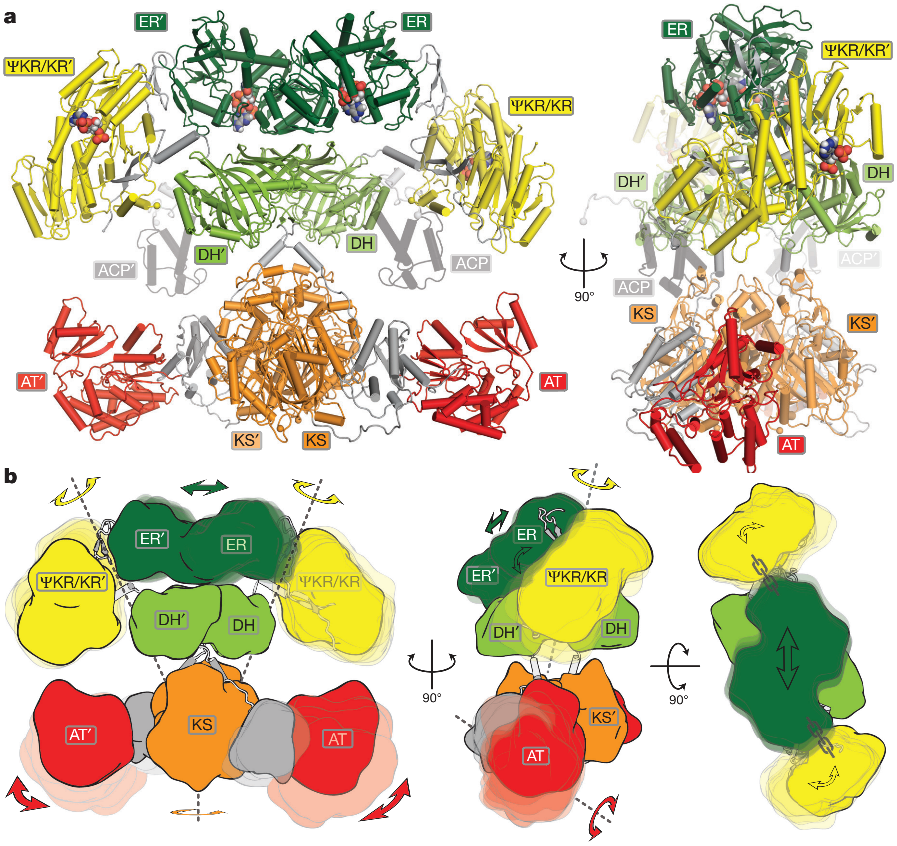


**Figure 4 | Hybrid model of a dynamic MAS dimer. a**, A hybrid MAS model was assembled by linking the condensing and modifying region structures. Central helical linkers connect the two regions without secondary interactions. The orientation around the linkage is presumably flexible and has been modelled according to the FAS structure. A homology model of mobile ACP is indicated (transparent grey) in a resting position without domain interactions. **b**, Conformational variability based on a comparison of 18 MAS modifying region chains and five homologous condensing region (Extended Data Fig. 6e, f) structures in combination with normal-mode analysis are shown. The lateral ER dimer motion on the DH platform is coupled to a rotation of both double-tethered ΨKR/KRs by up to 40.4° (Supplementary Video 1). The positions of the AT relative to KS in different condensing region structures correspond to a rotation of up to 43° between the most linear (MAS) and the most bent (human FAS) variant.

Our structural data reinforce the earlier hypothesis that modPKSs are fundamentally similar to non-colinear iPKSs such as MAS, and presumably evolved by kinetic coupling of modules[26]. Indeed, several modPKS modules act in an iterative mode as part of an assembly line (for example, BorA5 (ref. 27), AurA[28]). Other modPKS modules can be converted into a non-colinear mode of action by mutation, for example DEBS module 3 (ref. 29), or by isolation from their assembly line environment, for example PikAIII[30].

The analysis of the hybrid MAS structure depicts a unique PKS architecture. It agrees with previous biochemical and structural data on modPKSs fragments, with the exception of the monomeric state of some isolated ER domains[18] or in the domain-swapped crystal structure of the excised ER–ΨKR/KR di-domain of the fully reducing modPKS SpnB[19]. On the basis of the structure and the monomeric solution state of SpnB ER–ΨKR/KR, as well as shorter ER–KR linker in modPKSs, a divergent architecture of modPKSs modifying regions based on a dimeric DH arrangement with laterally positioned monomeric KRs and ERs was proposed[19]. On the contrary, the MAS modifying region retains a central dimeric ER as observed in FAS and in a fragment of the modPKS PpsC ER (Protein Data Bank accession number 1PQW) (Extended Data Fig. 3c). Importantly, MAS reveals a dynamic linker-based organization, which (in contrast to FAS) could also accommodate the typical range of ER–KR linker lengths (5–22 aa) observed in modPKSs (Extended Data Fig. 4) by slight adaptions of the ΨKR/KR position.

Models of PKS modifying regions based on SpnB ER–ΨKR/KR and MAS are clearly distinct on a macromolecular scale and can be experimentally distinguished via SAXS distance distributions. We selected two well-expressed modifying regions from modPKSs bi-modules, EryA of gammaproteobacterium HdN1 (GpEryA) and ‘Pks’ (Uniprot: Q3L885) from *M. smegmatis* (MsPks) for comparative SAXS analysis. Calculated SAXS curves for SpnB ER–ΨKR/KR and MAS-like models were compared with experimental SAXS data of MAS, GpEryA, and MsPks. The derived distance distributions closely match those calculated from a MAS-like model, but not those based on SpnB ER–ΨKR/KR (Extended Data Fig. 8). The SAXS analysis of GpEryA and MsPks clearly supports a wider relevance of the MAS architecture for modPKSs.

Our structural analysis not only provides detailed insights into MAS, a mycobacterial drug target, but also establishes a new template for the organization of PKSs modules. It reveals a unique, dynamic structure of the modifying region based on dimeric DH and ER domains and provides insights into conformational variability and coupling in fully reducing PKS modifying regions. The linker-based architecture supports modularity of the modifying region by requiring only the adaptation of variable linker regions for evolutionary domain shuffling. It thus rationalizes an important aspect of the notable success of the PKS architecture in the generation of chemical diversity. Our results highlight the relevance of matching linker–domain, rather than domain–domain, interactions in PKS engineering. They contribute to the fundamental understanding of PKS architecture, as well as to the functional dissection and re-engineering of related synthases including relevant drug targets and important producers of bioactive compounds.

**Online Content** Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

**Received 29 May 2015; accepted 12 January 2016.**

**Published online 14 March; corrected online 23 March 2016**

**(see full-text HTML version for details).**

1. Weissman, K. J. Uncovering the structures of modular polyketide synthases. *Nat. Prod. Rep.* **32,** 436–453 (2015).
2. Wong, F. T. & Khosla, C. Combinatorial biosynthesis of polyketides—a perspective. *Curr. Opin. Chem. Biol.* **16,** 117–123 (2012).
3. Hertweck, C. The biosynthetic logic of polyketide diversity. *Angew. Chem. Int. Edn Engl*. **48,** 4688–4716 (2009).
4. Busch, B. *et al.* Multifactorial control of iteration events in a modular polyketide assembly line. *Angew. Chem. Int. Edn Engl*. **52,** 5285–5289 (2013).
5. Moss, S. J., Martin, C. J. & Wilkinson, B. Loss of co-linearity by modular polyketide synthases: a mechanism for the evolution of chemical diversity. *Nat. Prod. Rep.* **21,** 575–593 (2004).
6. Maier, T., Leibundgut, M. & Ban, N. The crystal structure of a mammalian fatty acid synthase. *Science* **321,** 1315–1322 (2008).
7. Sirakova, T. D., Thirumala, A. K., Dubey, V. S., Sprecher, H. & Kolattukudy, P. E. The Mycobacterium tuberculosis pks2 gene encodes the synthase for the hepta- and octamethyl-branched fatty acids required for sulfolipid synthesis. *J. Biol. Chem.* **276,** 16833–16839 (2001).
8. Chopra, T. & Gokhale, R. S. Polyketide versatility in the biosynthesis of complex mycobacterial cell wall lipids. *Methods Enzymol.* **459,** 259–294 (2009).
9. Cambier, C. J. *et al.* Mycobacteria manipulate macrophage recruitment through coordinated use of membrane lipids. *Nature* **505,** 218–222 (2014).
10. Brignole, E. J., Smith, S. & Asturias, F. J. Conformational flexibility of metazoan fatty acid synthase enables catalysis. *Nature Struct. Mol. Biol.* **16,** 190–197 (2009).
11. Tang, Y., Kim, C. Y., Mathews, I. I., Cane, D. E. & Khosla, C. The 2.7-angstrom crystal structure of a 194-kDa homodimeric fragment of the 6-deoxyerythronolide B synthase. *Proc. Natl Acad. Sci. USA* **103,** 11124–11129 (2006).
12. Whicher, J. R. *et al.* Cyanobacterial polyketide synthase docking domains: a tool for engineering natural product biosynthesis. *Chem. Biol.* **20,** 1340–1351 (2013).
13. Dutta, S. *et al.* Structure of a modular polyketide synthase. *Nature* **510,** 512–517 (2014).
14. Edwards, A. L., Matsui, T., Weiss, T. M. & Khosla, C. Architectures of whole-module and bimodular proteins from the 6-deoxyerythronolide B synthase. *J. Mol. Biol.* **426,** 2229–2245 (2014).
15. Akey, D. L. *et al.* Crystal structures of dehydratase domains from the curacin polyketide biosynthetic pathway. *Structure* **18,** 94–105 (2010).
16. Keatinge-Clay, A. Crystal structure of the erythromycin polyketide synthase dehydratase. *J. Mol. Biol.* **384,** 941–953 (2008).
17. Gay, D., You, Y. O., Keatinge-Clay, A. & Cane, D. E. Structure and stereospecificity of the dehydratase domain from the terminal module of the rifamycin polyketide synthase. *Biochemistry* **52,** 8916–8928 (2013).
18. Khare, D. *et al.* Structural basis for cyclopropanation by a unique enoyl-acyl carrier protein reductase. *Structure* **23,** 2213–2223 (2015).
19. Zheng, J., Gay, D. C., Demeler, B., White, M. A. & Keatinge-Clay, A. T. Divergence of multimodular polyketide synthases revealed by a didomain structure. *Nature Chem. Biol.* **8,** 615–621 (2012).
20. Sippel, K. H., Vyas, N. K., Zhang, W., Sankaran, B. & Quiocho, F. A. Crystal structure of the human fatty acid synthase enoyl-acyl carrier protein-reductase domain complexed with triclosan reveals allosteric protein-protein interface inhibition. *J. Biol. Chem.* **289,** 33287–33295 (2014).
21. Keatinge-Clay, A. T. A tylosin ketoreductase reveals how chirality is determined in polyketides. *Chem. Biol.* **14,** 898–908 (2007).
22. Bonnett, S. A. *et al.* Structural and stereochemical analysis of a modular polyketide synthase ketoreductase domain required for the generation of a *cis*-alkene. *Chem. Biol.* **20,** 772–783 (2013).
23. Hardwicke, M. A. *et al.* A human fatty acid synthase inhibitor binds β-ketoacyl reductase in the keto-substrate site. *Nature Chem. Biol.* **10,** 774–779 (2014).
24. Bretschneider, T. *et al.* Vinylogous chain branching catalysed by a dedicated polyketide synthase module. *Nature* **502,** 124–128 (2013).
25. Whicher, J. R. *et al.* Structural rearrangements of a polyketide synthase module during its catalytic cycle. *Nature* **510,** 560–564 (2014).
26. Sugimoto, Y. *et al.* Freedom and constraint in engineered noncolinear polyketide assembly lines. *Chem. Biol.* **22,** 229–240 (2015).
27. Olano, C. *et al.* Biosynthesis of the angiogenesis inhibitor borrelidin by *Streptomyces parvulus* Tü4055: cluster analysis and assignment of functions. *Chem. Biol.* **11,** 87–97 (2004).
28. He, J. & Hertweck, C. Iteration as programmed event during polyketide assembly; molecular analysis of the aureothin biosynthesis gene cluster. *Chem. Biol.* **10,** 1225–1232 (2003).
29. Kapur, S. *et al.* Reprogramming a module of the 6-deoxyerythronolide B synthase for iterative chain elongation. *Proc. Natl Acad. Sci. USA* **109,** 4110–4115 (2012).
30. Beck, B. J., Aldrich, C. C., Fecik, R. A., Reynolds, K. A. & Sherman, D. H. Iterative chain elongation by a pikromycin monomodular polyketide synthase. *J. Am. Chem. Soc.* **125,** 4682–4683 (2003).

**Supplementary Information** is available in the online version of the paper.

**Acknowledgements** We acknowledge F. Widdel and J. Zedelius for providing gammaproteobacterium HdN1, P. Leadlay and L. Betancor for providing plasmid pETcoco-2A-L1SL2, and EMBL Heidelberg for providing the pETG-10A vector; J. Missimer and A. Menzel for support in SAXS data acquisition and raw data processing; T. Sharpe for analytical ultracentrifugation, A. Mazur for SAXS refinement, and M. Bertoni for support of the homology-based assignment of the oligomeric state of MAS KS–AT. Data were collected at beamlines PXI, PXIII, and cSAXS of PSI; we acknowledge support from the beamline teams. This work was supported by the Swiss National Science Foundation project grants 125357, 138262, 159696 and R'equip grant 145023. D.A.H. acknowledges a fellowship by the Werner-Siemens Foundation.

**Author Contributions** R.P.J. expressed, purified and crystallized MAS, obtained the crystal structure of the condensing region, collected SAXS data and cloned constructs. F.Z. cloned constructs and purified MAS, GpEryA and MsPks. D.A.H. purified MAS, optimized MAS crystallization, determined the structure of the isolated DH domains and the modifying region, collected SAXS data, analysed the data, performed homology modelling, cloned constructs, and wrote the manuscript. T.M. designed and guided research, analysed data, contributed to crystallographic analysis and wrote the manuscript.

**Author Information** Atomic coordinates and structure factors for the reported crystal structures have been deposited in the Protein Data Bank under accession numbers 5BP1, 5BP2, 5BP3, 5BP4. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to T.M. (timm.maier@unibas.ch).

## METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

**Cloning, expression, and purification.** *M. smegmatis* (ATCC 700084) was cultured according to ATCC recommendations. Cells were pelleted and washed with TE buffer. A cell pellet of gammaproteobacterium HdN1 was provided by J. Zedelius. Cells were resuspended in lysis buffer (0.1 M Tris pH 8.0, 0.2 M NaCl, 5 mM EDTA, 0.2 mg ml$^{-1}$ lysozyme), incubated for 6 h at 37 °C, subsequently supplemented with 0.5% SDS and 0.2 mg ml$^{-1}$ proteinase K, and incubated at 65 °C for 24 h. DNA was purified by phenol–chloroform extraction and dissolved in TE buffer. The MAS KS–AT constructs (Uniprot: A0R1E8, aa 1–884, 1–887, 1–892), MAS DH (A0R1E8, 884–1186), MAS DH–ΨKR–ER–KR (A0R1E8, 884–2020), and 'Pks' DH–ΨKR–ER–KR (Q3L885, 2450–3580) were cloned into pNIC28a-Bsa vectors; GpEryA DH–ΨKR–ER–KR (E1VID6, 2420–3575) constructs were cloned into a Gateway-compatible pETG-10A destination vector (provided by EMBL Heidelberg). MAS DH–ΨKR–ER–KR–ACP (A0R1E8, 884–2111) was cloned by codon-optimized gene synthesis of ACP (GenScript) and restriction cloning (BsrGI/HindIII) into pNIC28a-Bsa-MAS DH–ΨKR–ER–KR (884-2020). All constructs were designed as N-terminal tobacco etch virus (TEV) protease cleavable hexa-histidine (His6) fusion constructs and co-expressed with *Streptomyces* chaperonins[31] (pETcoco-2A-L1SL2 plasmid) in BL21(DE3) pRIL cells or for GpEryA in Rosetta(DE3) pLysS cells with a kanamycin-resistant version of pETcoco-2A-L1SL2. Cells were cultured in 2×YT media, supplemented with 0.5% glycerol, NPS (25 mM $(NH4)_2SO_4$, 50 mM $KH_2PO_4$, 50 mM $Na_2HPO_4$), kanamycin (100 μg ml$^{-1}$), chloramphenicol (34 μg ml$^{-1}$), and ampicillin (100 μg ml$^{-1}$). An expression culture (1.5 l) was inoculated (1:20), grown at 37 °C for 2 h, cooled to 20 °C, and induced with isopropyl-β-D-thiogalactopyranosid (0.1 mM) at an absorbance at 600 nm of 1.0. Cells were collected after 12 h by centrifugation (7,000*g*) and resuspended in lysis buffer (50 mM HEPES pH 7.4, 20 mM imidazole, 0.5 M NaCl, 5 mM $MgCl_2$, 10% glycerol (v/v), 2.5 mM β-mercaptoethanol), supplemented with protease inhibitors (200 μM phenylmethylsulfonyl fluoride, 20 μM bestatin, 4 μM E64, 2 μM pepstatin A, 20 μM phenantrolin, 2 μM phosphoramidon) as well as DNase, RNase, and lysozyme. Cells were placed on ice and lysed by sonication. The lysate was cleared by centrifugation (100,000*g*, 30 min) and the supernatant was loaded onto a 5 ml Ni-affinity column (GenScript) pre-equilibrated with lysis buffer. Unbound protein was eluted with four alternating wash cycles of five column volumes (CV) lysis buffer and HisA buffer (50 mM HEPES pH 7.4, 20 mM imidazole, 50 mM NaCl, 5 mM $MgCl_2$, 10% glycerol (v/v), 2.5 mM β-mercaptoethanol, inhibitors), until a stable baseline (A280) was reached. The sample was eluted with 2 CV HisB buffer (50 mM HEPES pH 7.4, 250 mM imidazole, 50 mM NaCl, 10% glycerol (v/v), 2.5 mM β-mercapto ethanol, inhibitors) and diluted (1:10) with AIC-A buffer (50 mM Tris-HCl pH 7.4, 20 mM KCl, 10% (v/v) glycerol, 2.5 mM β-mercaptoethanol). The sample was loaded on a 6.5 ml anion exchange column (PL-SAX 4,000 Å, 10 μm) and washed with 20 CV. The samples were eluted with a stepped gradient to 100% AIC-B (50 mM Tris-HCl pH 7.4, 1 M NaCl, 10% (v/v) glycerol, 2.5 mM β-mercaptoethanol). For DH–ΨKR–ER–KR the gradient was held at a conductivity of 15 mS cm$^{-1}$ until a stable baseline (A280) was obtained to elute *Streptomyces* chaperonins. DH–ΨKR–ER–KR eluted at 17–20 mS cm$^{-1}$. Pure fractions were pooled, supplemented with TEV protease (1 mg protease per 100 mg tagged protein), and incubated for 10 h at 4 °C. Uncleaved protein, as well as the cleaved His6-tag, was removed by passing the solution through a 5 ml orthogonal Ni-affinity column (GenScript). The flow-through was pooled, concentrated, and subjected to gel permeation chromatography (Superdex 200 16/60, GE Healthcare) using GPC buffer (20 mM HEPES pH 7.4, 250 mM NaCl, 5% glycerol (v/v), 5 mM dithiothreitol). Pure fractions were pooled, and monodispersity was monitored by dynamic light scattering at 1 mg ml$^{-1}$. Related purification protocols were applied to MAS DH–ΨKR–ER–KR–ACP, 'Pks' DH–ΨKR–ER–KR (HiTrap CaptoQ column), GpEryA DH–ΨKR–ER–KR (both no TEV protease cleavage and orthogonal Ni-affinity column), and MAS DH (no anion exchange chromatography).

**Crystallization.** All crystallization experiments were performed using a robotic setup applying the sitting drop vapour diffusion method.

MAS KS–AT bi-pyramidal crystals were grown at 4 °C by mixing 0.2 μl of protein in GPC buffer (38 mg ml$^{-1}$) with 0.2 μl reservoir solution (0.1 M MES/imidazole pH 6.5, 0.1 M $MgCl_2$, 0.1 M $CaCl_2$, 12.5% (v/v) polyethylene glycol 1000 (v/v), 7.5% (w/v) polyethylene glycol (PEG) 3350, 12.5% 2-methyl-2,4-pentanediol (MPD)). Crystals grew to a final size of 0.8 mm × 0.4 mm × 0.2 mm within 1 week and were flash-frozen in liquid nitrogen.

The MAS DH domain was crystallized in space group $P2_1$ at 18 °C by mixing 0.2 μl of protein in GPC buffer (38 mg ml$^{-1}$) with 0.1 μl reservoir solution (0.1 M bis-Tris pH 6.5, 0.2 M $MgCl_2$, 25% (v/v) PEG 3350) and grew to a final size of 0.4 mm × 0.2 mm × 0.1 mm within one week. Crystals in space group $P2_12_12$ appeared after 30 days at 18 °C by mixing 1 μl of protein in GPC buffer (38 mg ml$^{-1}$) with 2 μl reservoir solution (0.25 M di-sodium malonate, 24% (w/v) PEG 3350) and grew to a final size of 1 mm × 0.4 mm × 0.3 mm. Before harvesting, all crystals of MAS DH were cryoprotected (25% (v/v) ethylene glycol) and flash-frozen in liquid nitrogen.

Needle-shaped crystals of MAS DH–ΨKR–ER–KR were obtained by mixing protein solution at 18.4 mg ml$^{-1}$ (GPC buffer, 1.5 mM NADP$^+$) and reservoir solution (0.03 M $MgCl_2$, 0.03 M $CaCl_2$, 20% ethylene glycol, 10% PEG 8000, 0.1 M MES/imidazole pH 6.5) at 4 °C. Crystallization was optimized by exchanging PEG 8000 by PEG 3350, decreasing the PEG 3350 concentration to 7–13% (w/v), and by carefully monitored microseeding. Subsequent optimization was performed using automated robotic setup and seeding at 4 °C. Final crystals (1.0 mm × 0.3 mm × 0.2 mm) were obtained after mixing 1 μl protein (20.3 mg ml$^{-1}$ in GPC buffer incl. 1.5 mM NADP$^+$) with 1 μl of reservoir solution (5.25% (w/v) PEG 3350, 20% (v/v) ethylene glycol, 0.1 M MES pH 7.0, 52 mM $MgCl_2$, 52 mM $CaCl_2$) and 0.2 μl seed stock. Diffraction properties were optimized by crystal dehydration: over a period of 4 h, crystals were transferred to a dehydration solution (0.05 M MES pH 7.0, 25% ethylene glycol, 25% PEG 3350, 56 mM $MgCl_2$, 56 mM $CaCl_2$, 1.5 mM NADP$^+$) by a step-wise exchange of the drop solution. All crystals were harvested and plunge-frozen in liquid nitrogen. Integrity of the protein in final crystals was examined by denaturing polyacrylamide gel electrophoresis.

**Data collection and structure determination.** All data sets were collected at the Swiss Light Source (Villigen, Switzerland) at a temperature of 100 K. Data sets of DH crystals were collected at beamline X06DA ($P2_1$, $\lambda = 0.999870$ Å, $T = 100$ K; $P2_12_12$, $\lambda = 0.97626$ Å). All other data sets were collected at beamline X06SA (KS–AT, $\lambda = 0.97940$ Å; DH–ΨKR–ER–KR, $\lambda = 0.97626$ Å). Data reduction was performed using XDS[32] and XSCALE[32], and data sets were analysed with phenix.xtriage[33]. All structures were solved with PHASER[34] using molecular replacement.

Crystals of all KS–AT di-domain variants of MAS are isomorphic in space group $P4_12_12$. The KS and AT domains of DEBS $KS_5$-$AT_5$ (ref. 11) were used as molecular replacement templates and initial rebuilding was achieved by BUCCANEER[35]. All three crystal structures were virtually identical except for the identity of the last ordered C-terminal residue. The construct with the most extended C terminus (1–892) revealed aa 887 as the last ordered residue, overlapping in sequence with the modifying region. Thus we continued refinement only for crystals of this variant (aa 1–892) with unit cell constants of $a = 77.5$ Å, $b = 77.5$ Å, $c = 371.2$ Å, and a solvent content of 56%. A final model was obtained after iterative cycles of real space model building in COOT[36] and TLS refinement in Phenix[33], and was refined to $R_{\mathrm{work}}/R_{\mathrm{free}}$ values of 0.21/0.23 at 2.3 Å resolution with excellent geometry (Ramachandran favoured/outliers: 97.8%/0.2%) (Extended Data Table 1a).

Crystals of the DH domain of MAS belong to space group $P2_1$ ($a = 59.7$ Å, $b = 162.4$ Å, $c = 66.6$ Å, $\beta = 91.4°$) and $P2_12_12$ ($a = 67.1$ Å, $b = 162.2$ Å, $c = 59.5$ Å) with a solvent content of 49% and 51%, respectively. A molecular replacement search model was based on CurK DH[15]. Initial maps were improved by density modification and NCS averaging with PARROT[37], followed by automated rebuilding with BUCCANEER[35]. Final models were obtained after iterative cycles of model building in COOT[36], and refinement in BUSTER[38] ($P2_1$) and Phenix[33] ($P2_12_12$), yielding excellent geometry (Ramachandran favoured/outliers: $P2_1 = 98.2\%/0.0\%$; $P2_12_12 = 98.2\%/0.2\%$) and $R_{\mathrm{work}}/R_{\mathrm{free}}$ values of 0.18/0.20 ($P2_1$) and 0.15/0.18 ($P2_12_12$) (Extended Data Table 1a).

Crystals of MAS DH–ΨKR–ER–KR in space group $P1$ ($a = 151.4$ Å, $b = 190.4$ Å, $c = 270.8$ Å, $\alpha = 95.6°$, $\beta = 91.9°$, $\gamma = 103.7°$) diffracted to a maximum resolution of 3.75 Å. The asymmetric unit contained 18 protomers in 9 dimers with 20,502 amino acids and a molecular mass of 2.2 MDa at 65% solvent content. Data were collected at four different positions of a single crystal and combined to obtain a complete high-quality data set. The resolution cutoff was determined by $CC_{1/2}$ criterion[39]. Self-rotation functions revealed rotational NCS and the native Patterson function indicated translational NCS.

Initially, a partial molecular replacement solution was obtained for the ER dimer using the ER domain of porcine FAS (pFAS)[6]. Other known structures of homologous domains did not provide efficient search models. The structure of the isolated MAS DH domain, determined here independently, yielded equivalent solutions in agreement with the pFAS ER based solution. For final structure determination, both models where used in subsequent rounds of molecular replacement. Start models for building further regions were generated by homology modelling using Swiss Model[40]. To allow unbiased refinement in real and reciprocal space, phenix.reflection_tools[33] was used to define a thin-resolution shell-based test set[41], and test set reflections were excluded from calculating maps, which were used for real-space refinement. Initial refinement cycles included rigid body refinement and restrained refinement. The impact of various low-resolution restraint formulations on refinement was tested carefully. Local NCS is particularly well suited for MAS DH–ΨKR–ER–KR intermediate resolution refinement because of the high degree of NCS and the fact that using local NCS restrains avoids any external standard restraints based on assumptions on secondary structure or homologous peptide structures. Thus local structural similarity restraints[42] were combined only with reference model restraints to the authentic DH domain structure using autopruning in BUSTER[38]. After every round of refinement, bias-reduced, solvent flattened, and NCS-averaged maps were calculated using DM[43] without applying phase combination. Sharpened NCS-average maps were generated by applying a sharpening *B*-factor to the structure factor amplitudes before averaging. Initially, real-space rigid body fitting of individual secondary structure elements was applied for instances of every domain type (DH, ΨKR, $ER_{NB}$, $ER_{SB}$, KR) followed by symmetry expansion and rigid body fitting for entire domains. Best-defined regions of the electron density maps were used for rebuilding of every domain type using Coot[36] and O[44], respectively, symmetry expanded, and recombined into 18 chains. At this point, unambiguous difference electron density indicated the connecting linkers, which were manually built into the maps and refined without symmetry expansion (Extended Data Fig. 2d, e). Later refinement cycles included TLS refinement, using one group per domain and linker, individual *B*-factor refinement, and automated weight factor determination. During rebuilding, B-sharpening, NCS average, and density modification, as well as feature enhanced maps[33] were used. Overall, the use of 18-fold-domain-wise NCS averaging results in highly accurate and unbiased phase determination, irrespective of details of the atomic model. The combined use of NCS-averaging and *B*-factor sharpening led to an exceptional map quality typical for maps at considerably higher resolution (Extended Data Fig. 2f). Bound NADP$^+$ cofactors were added for final refinement cycles. NADP$^+$ is well ordered in the ER domain, while the nicotinamide moieties are disordered in the KR domains and were not included in the final model. A total of five KR and four ΨKR domains, which lack stabilization by crystal contacts, were either disordered or present in multiple orientations, and not included in the final model, despite substantial positive difference density. The KR domain in chain L shows a considerably more tilted orientation as observed in all other instances of the KR domains, which, however, agrees with the identified hinge regions. A single model was placed for this domain, which achieved the largest improvement of *R*-factors and was characterized by the lowest *B*-factors after refinement, although a secondary alternative conformation might be present. The refinement of the final model (excluding disordered regions (chains): 883–895 (E–R), 1206–1213, 1283–1287, 1948–1960, ΨKR(I/L/O/Q–R), KR(I/O/Q–R)) was completed with $R_{\text{work}}/R_{\text{free}}$ values of 0.23/0.24 and very good geometry for the resolution range (Ramachandran favoured/outliers: 91.6%/1.8%).

**Analytical ultracentrifugation.** To determine oligomeric states in solution, sedimentation equilibrium analytical ultracentrifugation experiments were performed for MAS DH–ΨKR–ER–KR and MAS KS–AT. Columns (140 μl) containing proteins at concentrations of 3.5–4.5 mg ml$^{-1}$ in GPC buffer were subjected to centrifugation at 4,800 and 7,800 r.p.m. in the Beckman An-50 Ti rotor, corresponding to 1,800*g* and 4,760*g* at the radial midpoint of the solution column, at 12 °C, with detection by radial absorbance scanning at 305 nm. At each speed, centrifugation was allowed to proceed until sedimentation equilibrium was attained, as judged by pairwise comparison of scans using the approach to equilibrium function in SEDFIT (https://sedfitsedphat.nibib.nih.gov). Buffer density (1.0277 g ml$^{-1}$) and viscosity (1.5306 centipoise) were measured at 12 °C using an Anton Paar DMA4500M densitometer and an AMVn viscometer, respectively. Molar extinction coefficients at 305 nm were calculated for each protein from the ratio of observed absorbance at various wavelengths in spectra at different dilutions and calculated molar extinction coefficients. The partial specific volume for each protein was calculated from sequence in SEDFIT. The radial absorbance scans at equilibrium for the two speeds were globally fitted to the 'single species of interacting system' mode in SEDPHAT[45] to determine the apparent molecular mass of the protein in solution. If the obtained molecular mass was intermediate between the value expected for a monomer and a dimer, the data were globally fitted to the monomer–dimer association model in SEDPHAT, with the molecular mass of the monomer fixed to the value calculated from the sequence. In both cases data were fitted using a fixed meniscus position, a floating bottom position, mass conservation constraints, a floating baseline and fitting radially independent noise components. Confidence intervals on single-species masses or dissociation constants were obtained by the Monte-Carlo method implemented in SEDPHAT.

**SAXS.** SAXS data were collected at beamline X12SA of Swiss Light Source. Samples were dialysed into GPC buffer, diluted to concentrations between 3 and 10 mg ml$^{-1}$and centrifuged at 13,000*g* and 8 °C until measurement. Glass capillaries (1 mm inner diameter) were mounted on a temperature-controlled holder at 12 °C. Data collection was performed using a Pilatus 2M detector at a distance of 2.14 m and a wavelength of 1.000 Å. Data were collected in eight repetitive scans each including ten 40 ms acquisitions at ten capillary positions yielding a total of 800 frames per buffer and protein, respectively. Frames with artefacts for example, from air bubbles, were identified using Swiss Light Source/PSI software (SAXS_inspect2) and excluded from the data sets. Radial averages were calculated and exported using beamline software for scattering vectors from 0.005 to 0.7 Å$^{-1}$ defined as $q = 4\pi/\lambda \sin\theta$. Scattering curves were averaged using DATAVER[46]; buffer profiles were subtracted using DATOP[46]. Scaling factors and *P* values of a Student's *t*-test were analysed using DATMERGE[46] and DATCMP[46], respectively. Later frames were affected by increasing radiation damage and were excluded from further processing. Final scattering curves for each sample concentration were thus obtained from 300 individual profiles. The radius of gyration ($R_g$) and zero angle intensity ($I(0)$) was calculated from the Guinier approximation using AUTORG[46] and is consistent with values obtained from atomic distance distributions $p(r)$ using DATGNOM[46] (Extended Data Table 1b). Scattering profiles at different concentrations were only combined if a noise reduction at medium and high scattering vectors could be obtained.

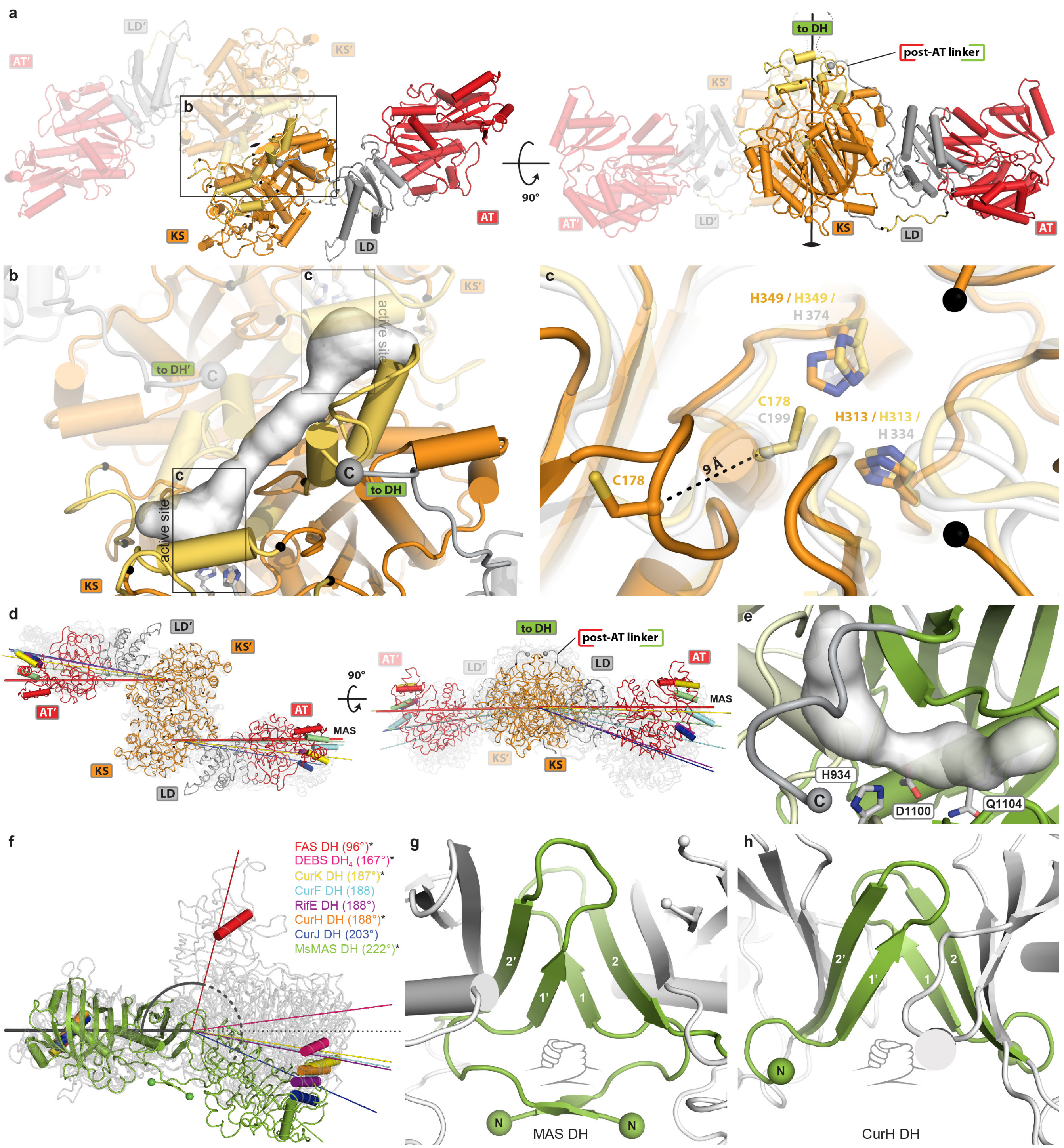
Modifying regions bear an intrinsic flexibility, which requires a flexible fitting approach to sample the full conformational space of the structures. Some approaches for flexible SAXS fitting have been described[47,48], but none was able to refine an individual structure while maintaining two-fold symmetry. Therefore, we combined dynamic elastic network restraints from CNS[49] with SAXS-target refinement and two-fold symmetry averaging in XPLOR-NIH[50] for the refinement of individual structures by simulated annealing. SAXS scattering curves of atomic models, fits with experimental data, and distance distributions were calculated using CRYSOL[46] and DATGNOM[46]. All SAXS curves were plotted using Python Matplotlib.

To compare calculated and experimental SAXS scattering curves, three models for the architecture of modifying regions were generated on the basis of the crystal structure of the domain-swapped SpnB fragment (ER–KR/ΨKR). The first model was obtained according to the original publication[19] by superposing the monomeric ER–KR/ΨKR domain on the KR domain of pFAS[6]. A linear homology model of SpnB DH[40] was placed into the position of pFAS DH and the domain swap in SpnB ER–KR/ΨKR was replaced with the corresponding region from DEBS $KR_1$ (ref. 51). The second model was constructed in the same way via a superposition on MAS KR. The relative domain arrangement of SpnB ER–KR/ΨKR was not altered in these two models, only the domain swap was corrected. The third, more generalized modPKS model was constructed to verify if shorter ER–KR linkers are in contradiction with the architecture of MAS. As a representative for short ER–KR linkers, the structure of SpnB ER–KR/ΨKR (6 aa) and the corresponding DH homology model were modelled as individual domains on MAS, while the linear DH dimer was maintained. ΨKR–ER linkers could be readily reconnected and regularized, whereas the ER–KR linker required a tilt of the ΨKR/KR domain. The tilt maintained a reasonable distance between the C terminus of the DH and the N terminus of the ΨKR domain, and yielded a linker architecture of a modPKS in agreement with MAS without stable direct interdomain contacts. SAXS curves and distance distributions were calculated for all models and compared with experimental SAXS scattering curves of MAS and two modPKS modifying regions with short ER–KR linkers (GpEryA, 9 aa; MsPks, 8 aa).

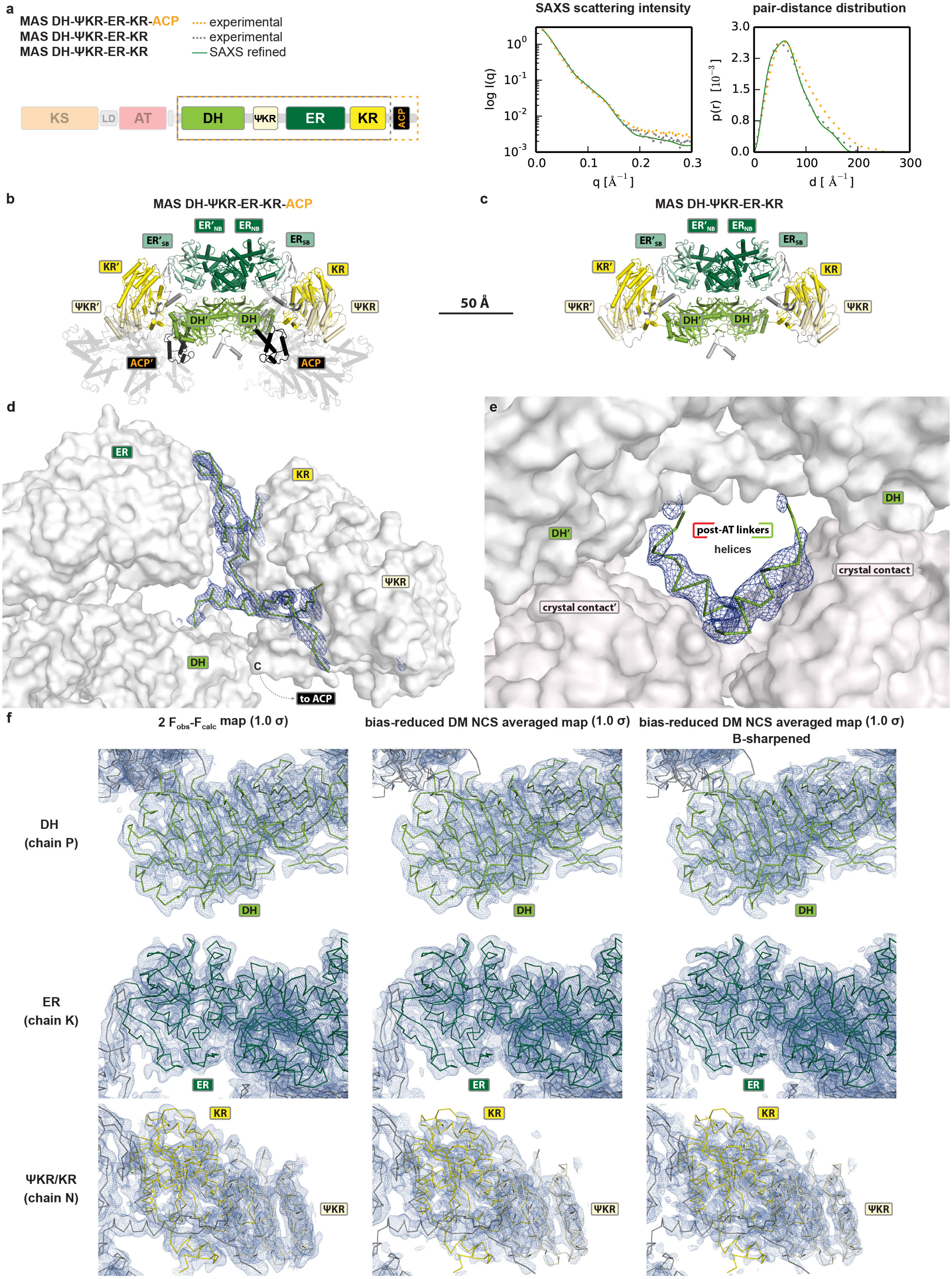
**Structure analysis and visualization.** Related structures were identified using PDBeFold[52] and interfaces were analysed using QtPISA[53]. Transformations and coordinate manipulations were performed using CCP4 (ref. 54) tools, MODTRAFO (T. Schirmer; http://www.biozentrum.unibas.ch), and MOLEMAN[55]. The automated Oligo algorithm[56] as implemented in Swiss Model unambiguously detected and predicted a single mode of dimerization of MAS KS–AT based on sequence homology. Initially, the dimeric form of KS–AT was assembled by least-squares fitting of secondary structure elements on DEBS $KS_5$ (ref. 11). Then, all residues in a radius of 7.5 Å to the dimer interface were deleted and multi-template homology modelling using modeller 9.15 (ref. 57) was used to construct a full-length dimeric homology model based on 20 homodimeric PKSs/FASs KS structures and the interface deleted MAS KS–AT structure. Remodelled regions (excluding all crystallographically defined regions beyond the radial cutoff) were geometry minimized using phenix.geometry_minimization[33]. The position where the post-AT linker becomes disordered was located by crystallization of KS–AT di-domains with three different linker lengths (1–884, 1–887, 1–892). Normal mode analyses were performed using the Bio3D[58] library in 'R'. Hinge bending analysis was performed by pre-aligning all structures to a reference substructure using LSQKAB[59], followed by a MODTRAFO (T. Schirmer; http://www.biozentrum.unibas.ch) analysis of the moving substructure. Principle screw axes were determined by averaging the direction vectors of the screw axes using Python Numpy and locating a central hinge point from the position of all screw axes. Active-site distances were calculated using BIOPYTHON[60]. All axes were visualized using PYMOL[61]. Interdomain angles of DH dimers were calculated by pre-aligning all DH dimers to one DH domain of MAS DH, followed by calculating the angle between the first principle component vector of the secondary structure elements of both domains. The angles were visualized using PYMOL[61]. Bias-removal for $F_{obs} - F_{calc}$ omit maps was achieved by applying a random perturbation to coordinates (Δ0.2 Å) and *B*-factors (Δ20% of the mean overall *B*-factor) using MOLEMAN2 (ref. 55) before refinement. Figures, videos, and active-site tunnels were generated using PYMOL[61], LSQMAN[62], and CAVER 3.0 (ref. 63).

**Sequence analysis.** Fifty-five sequences containing fully reducing modifying regions were selected from FASs, fiPKSs, Msl-, one *trans*-AT, and 36 modPKSs modules. Structure-based sequence alignments of all PKSs/FASs type I domain structures were generated using PDBefold[52] and used as reference for the alignment of individual domains using ClustalW2 (ref. 64). Linkers were aligned without reference, assembled with the individual domain alignments, and manually corrected in Geneious version 7.1.7 (ref. 65). Phylogenetic trees were generated using the neighbouring joining algorithm in Geneious version 7.1.7 (ref. 65).

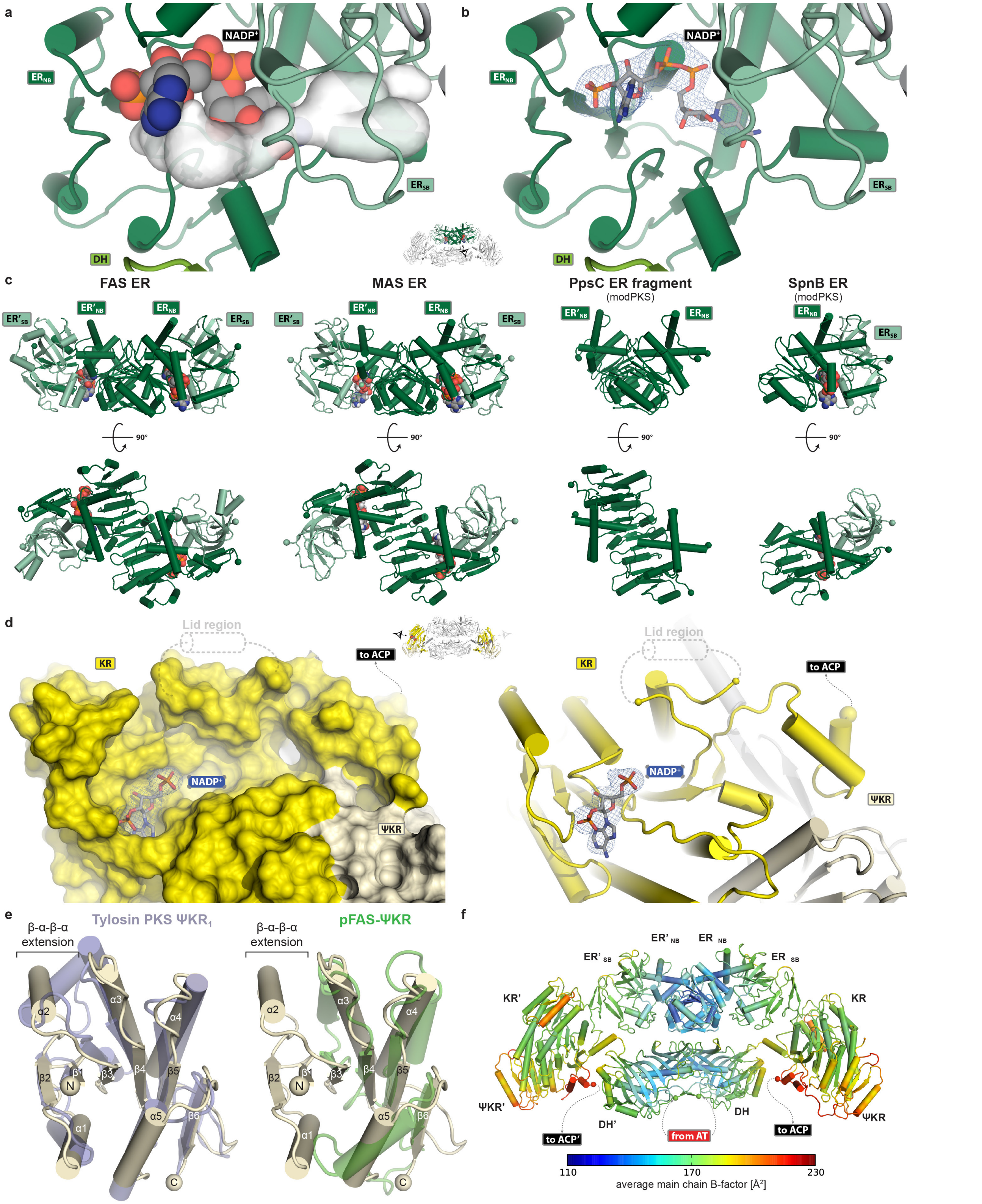
31. Betancor, L., Fernández, M. J., Weissman, K. J. & Leadlay, P. F. Improved catalytic activity of a purified multienzyme from a modular polyketide synthase after coexpression with *Streptomyces* chaperonins in *Escherichia coli*. *ChemBioChem* **9,** 2962–2966 (2008).
32. Kabsch, W. XDS. *Acta Crystallogr. D* **66,** 125–132 (2010).
33. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
34. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
35. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D* **62,** 1002–1011 (2006).
36. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
37. Cowtan, K. Recent developments in classical density modification. *Acta Crystallogr. D* **66,** 470–478 (2010).
38. Bricogne, G. B. E. *et al.* BUSTER version 2.10.2 (Global Phasing, 2011).
39. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science* **336,** 1030–1033 (2012).
40. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* **31,** 3381–3385 (2003).
41. Fabiola, F., Korostelev, A. & Chapman, M. S. Bias in cross-validated free R factors: mitigation of the effects of non-crystallographic symmetry. *Acta Crystallogr. D* **62,** 227–238 (2006).
42. Smart, O. S. *et al.* Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr. D* **68,** 368–380 (2012).
43. Cowtan, K. An automated procedure for phase improvement by density modification. *Joint CCP4 ESF-EACBM Newslett. Protein Crystallogr.* **31,** 34–38 (1994).
44. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
45. Vistica, J. *et al.* Sedimentation equilibrium analysis of protein interactions with global implicit mass conservation constraints and systematic noise decomposition. *Anal. Biochem.* **326,** 234–256 (2004).
46. Petoukhov, M. V. *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45,** 342–350 (2012).
47. Pelikan, M., Hura, G. L. & Hammel, M. Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys.* **28,** 174–189 (2009).
48. Zheng, W. & Tekpinar, M. Accurate flexible fitting of high-resolution protein structures to small-angle x-ray scattering data using a coarse-grained model with implicit hydration shell. *Biophys. J.* **101,** 2981–2991 (2011).
49. Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
50. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160,** 65–73 (2003).
51. Keatinge-Clay, A. T. & Stroud, R. M. The structure of a ketoreductase determines the organization of the β-carbon processing enzymes of modular polyketide synthases. *Structure* **14,** 737–748 (2006).
52. Krissinel, E. & Henrick, K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr. D* **60,** 2256–2268 (2004).
53. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372,** 774–797 (2007).
54. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).
55. Kleywegt, G. J. Validation of protein models from Cα coordinates alone. *J. Mol. Biol.* **273,** 371–376 (1997).
56. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42,** W252–W258 (2014).
57. Šali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).
58. Skjærven, L., Yao, X. Q., Scarabelli, G. & Grant, B. J. Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics* **15,** 399 (2014).
59. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **32,** 922–923 (1976).
60. Cock, P. J. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).
61. Schrodinger, L. The PyMOL Molecular Graphics System, version 1.7.0.3 (2010).
62. Kleywegt, G. J. Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallogr. D* **52,** 842–857 (1996).
63. Medek, P. B. P. & Sochor, J. Computation of tunnels in protein molecules using Delaunay triangulation. *J. WSCG* **15,** 107–114 (2007).
64. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
65. Kearse, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28,** 1647–1649 (2012).
66. Pappenberger, G. *et al.* Structure of the human fatty acid synthase KS-MAT didomain as a framework for inhibitor design. *J. Mol. Biol.* **397,** 508–519 (2010).
67. Tang, Y., Chen, A. Y., Kim, C. Y., Cane, D. E. & Khosla, C. Structural and mechanistic analysis of protein interactions in module 3 of the 6-deoxyerythronolide B synthase. *Chem. Biol.* **14,** 931–943 (2007).



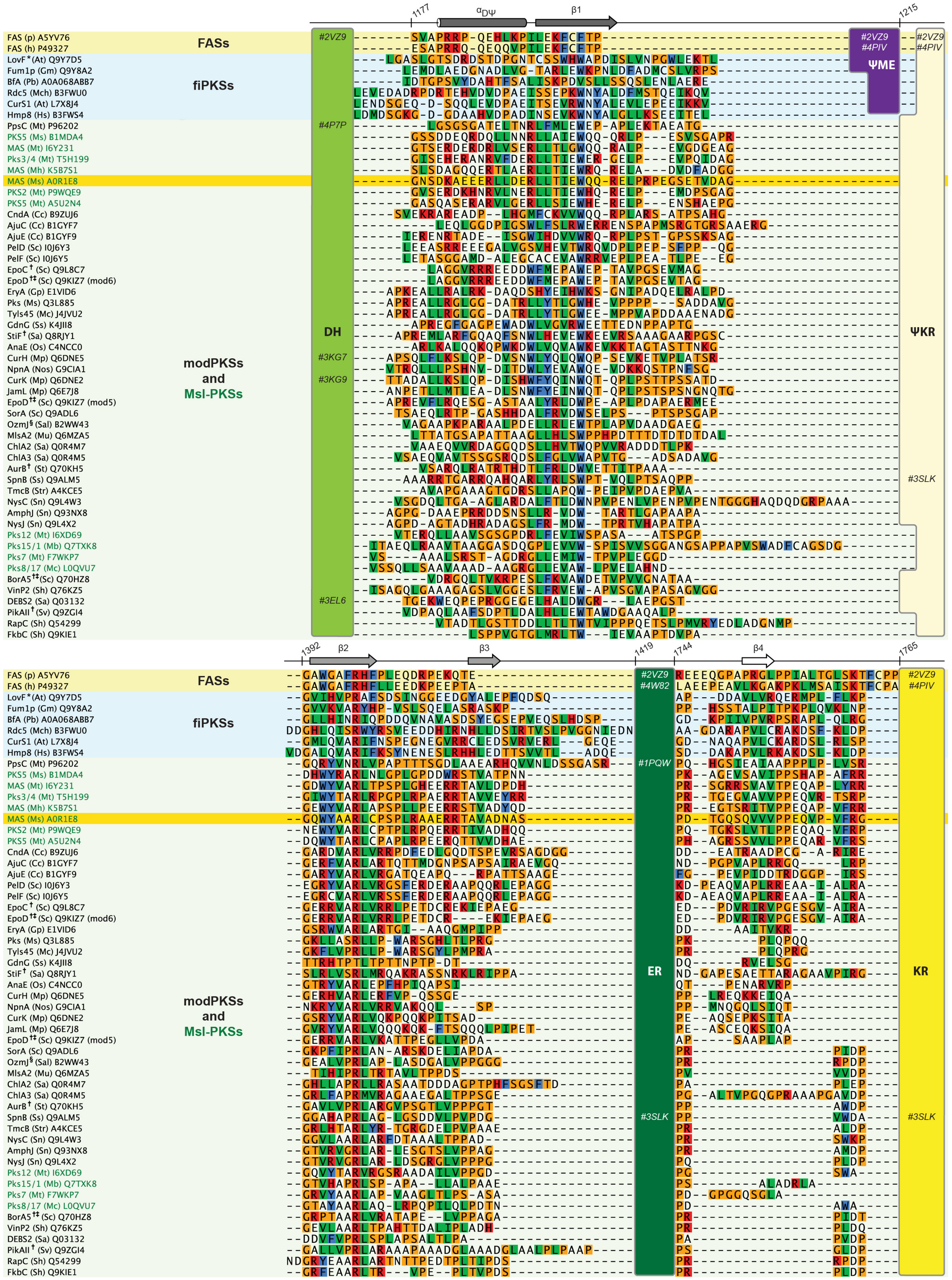
**Extended Data Figure 1 | Reconstruction of the dimeric KS–AT di-domain and DH dimer organization. a**, The condensing region dimer was reconstructed by least-square fitting on DEBS KS$_5$ (ref. 11) and multi-template homology modelling of disordered segments and the active-site loop (gold). Termini of the remodelled segments are indicated by black spheres. A pseudo-continuous β-sheet is formed across the dimer interface. The post-AT linker terminates close to the dimer axis. **b**, Close-up view on the reconstructed KS dimer with an active-site tunnel spanning both protomers (white), which is enclosed by four remodelled segments (gold). **c**, The active-site loop containing the catalytic Cys178 is dislocated in the monomeric (orange) form of MAS KS–AT, whereas the active-site His313 and His349 occupy the same position as in the dimeric DEBS KS$_5$-AT$_5$ structure (white-transparent). The canonical conformation of Cys178 observed in dimeric KS domains is restored in the dimeric KS–AT model (gold-transparent) **d**, MAS KS–AT (coloured, red line) reveals the most linear overall structure (right) of all PKSs/FAS condensing region structures[6,11,12,66,67] (corresponding to Extended Data 6e, f). **e**, The DH active-site residues are located at the interface of the two hot-dog folds (light and dark green; active-site tunnel in white). **f**, Interdomain angles in DH dimers[6,15–17]. Dimers were superposed onto one protomer (left) of MAS, and the angles between two protomers are compared. For clarity, only MAS DH is shown in green, for other DH domains only one equivalent helix is highlighted in colour. The FAS pseudo-dimeric DH domains (red helix) adopt a V-shaped structure (interdomain angle 96°), while PKS DH dimers (various colours) are almost linear (167–203°). The MAS DH dimer (green) is bent to the opposite direction relative to FAS, and exhibits the largest interdomain angle (222°) (asterisks indicate DHs that are part of fully reducing modifying regions). **g**, **h**, Dimer interface of MAS DH (**g**) and dimer interface of the isolated DH of the CurH[15] modPKS (**h**). Dimerization of MAS and CurH DH are mediated by ‘handshake’ interactions of the N-terminal hot-dog folds. In MAS DH, an N-terminal β-strand extension further contributes to dimerization.



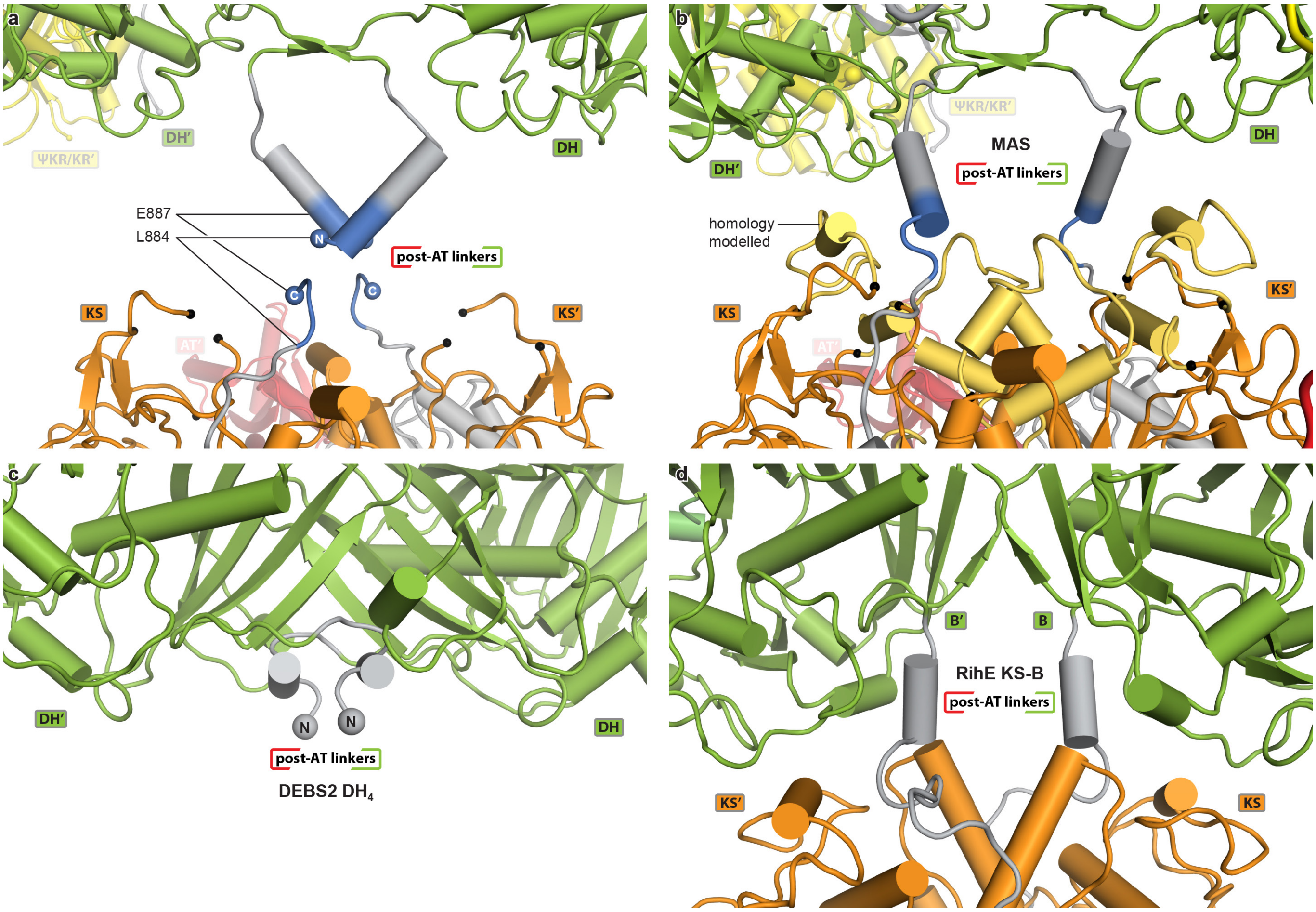
**Extended Data Figure 2 | Effect of ACP deletion and electron density maps of the MAS modifying region crystal structure. a**, SAXS experiments reveal conserved scattering profiles for the modifying region with ACP (dotted orange) and without ACP (dotted green), which resemble the scattering curve of the SAXS-refined X-ray structure (green). **b**, **c**, The experimentally determined interatomic distance distributions are in agreement with the maximum extends of the modifying domain with (**b**) and without (**c**) ACP, 250 Å and 201 Å, respectively. In **b** a set of plausible ACP positions is shown (transparent), on the basis of the length of the KR–ACP linker. **d**, Unbiased $F_{\mathrm{obs}} - F_{\mathrm{calc}}$ omit difference map of the modifying region linkers in chain B (contoured at 2.5$\sigma$) is shown. **e**, Unbiased $F_{\mathrm{obs}} - F_{\mathrm{calc}}$ omit difference map of the post-AT linker helices in chain A and B (contoured at 2.5$\sigma$); the helices could be modelled because of stabilizing crystal contacts. **f**, Electron density maps covering the three different domain types as indicated (left: $2F_{\mathrm{obs}} - F_{\mathrm{calc}}$ at 1.0$\sigma$; middle: bias-reduced density modified NCS average map at 1.0$\sigma$; right: bias-reduced density modified NCS average map at 1.0$\sigma$, with additional details revealed by applying a B-sharpening factor of − 80 Å$^2$).



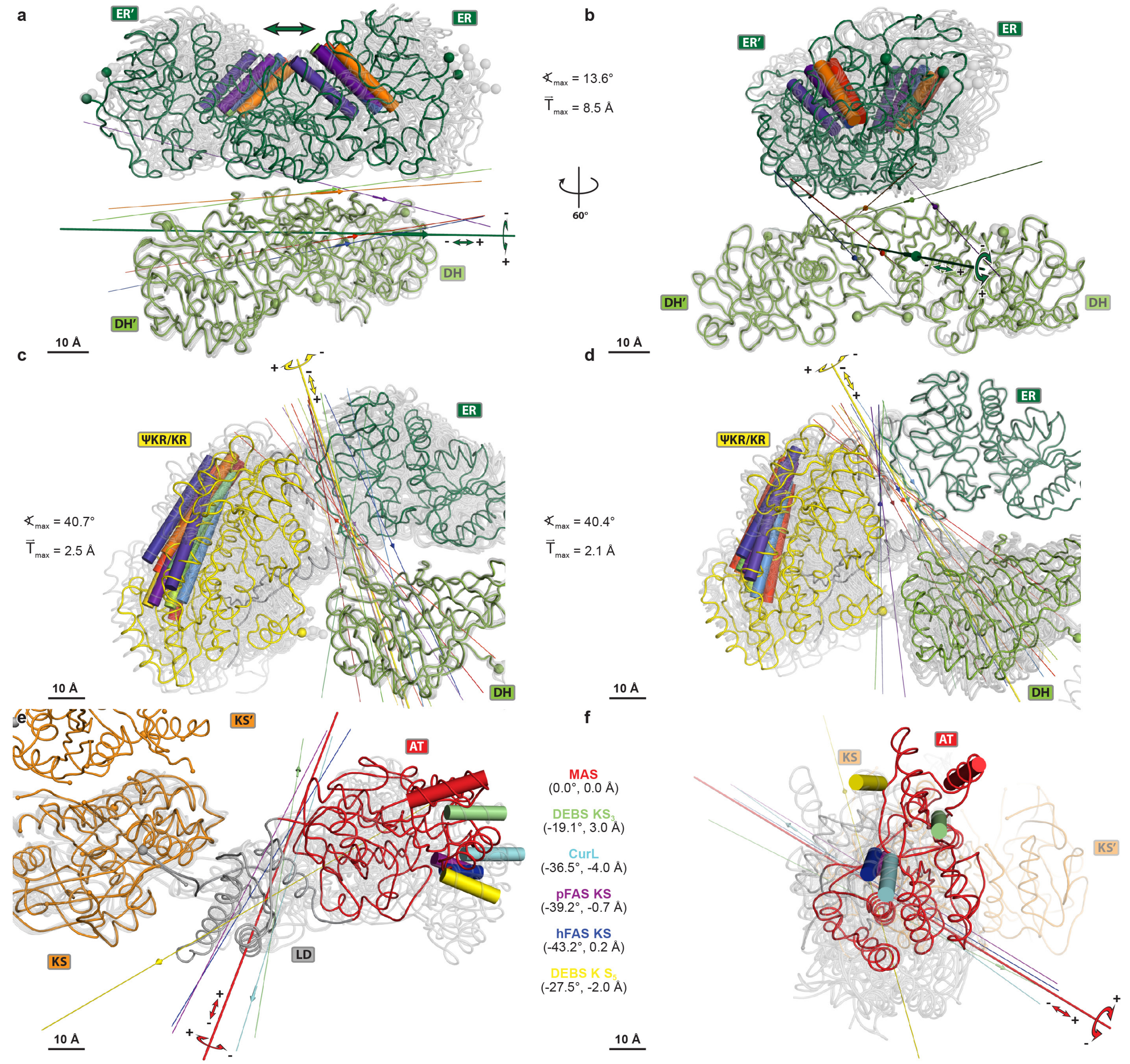
**Extended Data Figure 3 | Active site and structural comparison of the MAS ER and ΨKR/KR domains. a**, The MAS ER active-site tunnel (white) is lined by an NADP$^+$ cofactor. **b**, An $F_{\mathrm{obs}} - F_{\mathrm{calc}}$ shaked omit map (contoured at 3.0$\sigma$) is shown for the NADP$^+$ cofactor in chain J. **c**, The ER domains of FAS[6], MAS, and the modPKS PpsC dimerize via continuous β-sheet formation between the nucleotide binding subdomains (ER$_{\mathrm{NB}}$), whereas the SpnB ER was crystallized as monomer and represents a group of isolated ER domains[18,19]. **d**, The active site of ΨKR/KR locates to an elongated surface groove, which partly extends to the ΨKR domain and is presumably closed upon ligand binding by a disordered lid region (aa 1948–1960). An $F_{\mathrm{obs}} - F_{\mathrm{calc}}$ omit map (contoured at 3.0$\sigma$) is shown for the partly ordered NADP$^+$ cofactor. Left: surface; right: cartoon representation. **e**, MAS (pale yellow) features an N-terminal $\beta_1$–$\alpha_1$–$\beta_2$–$\alpha_2$ extension of the ΨKR Rossmann-fold, which is commonly found in PKSs (violet: tylosin PKS ΨKR$_1$ (ref. 21)), but absent in FASs (green: porcine FAS (pFAS) ΨKR[6]). Secondary structure labels refer to MAS ΨKR. **f**, Average main chain *B*-factors across all chains reveal distally increasing flexibility with highest *B*-factors for the ΨKR domain, in particular its β–α–β–α extension, and the C-terminal ACP anchor.



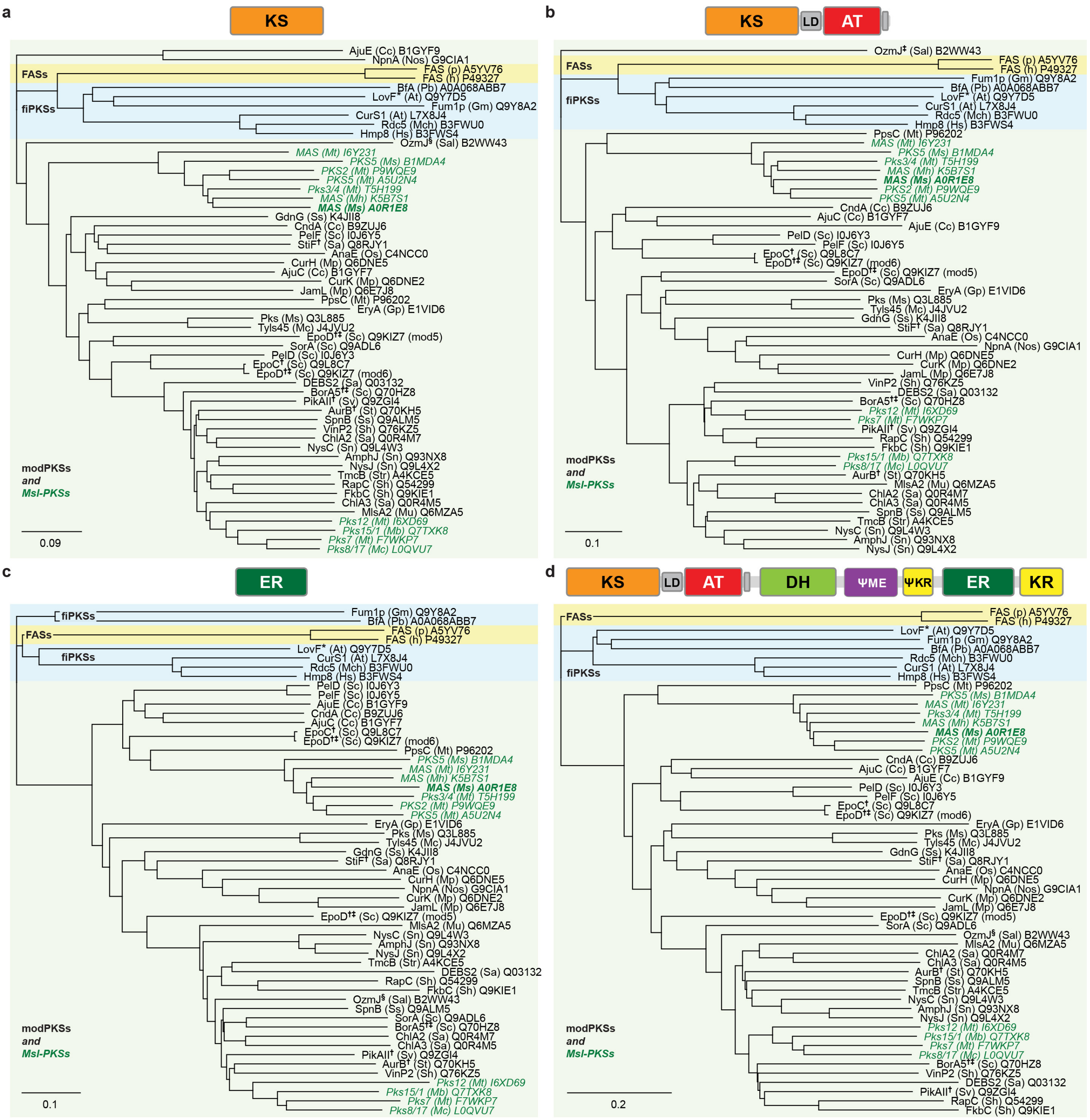
**Extended Data Figure 4 | Alignment of linker regions of 55 fully reducing modifying regions of PKSs and FASs.** The alignment reveals sequence conservation of the β-sheet B1 (β1 and β2), which is inserted in a surface groove of the ΨKR/KR domain. In MAS, strands β3 and β4 form the second antiparallel β-sheet B2. The ER–KR linker is considerably shorter in a subgroup of modPKSs. Sequence numbers and secondary structure elements correspond to *M. smegmatis* MAS (MAS (Ms) highlighted in orange). All modules are labelled as protein name (organism abbreviation) Uniprot number. Modules of Msl-PKSs (green text), modPKSs (light green), fiPKSs (blue), and FASs (yellow) are grouped by phylogeny (for details and colour coding see Extended Data Fig. 7). Protein Data Bank accession numbers are indicated in the boxes representing the corresponding domains. Amino acids are shown in clustal colours. (*Diketide synthase; †PKS cluster contains non-colinear iterative modules; ‡modular non-colinear iPKS module; §*trans*-AT PKS.)



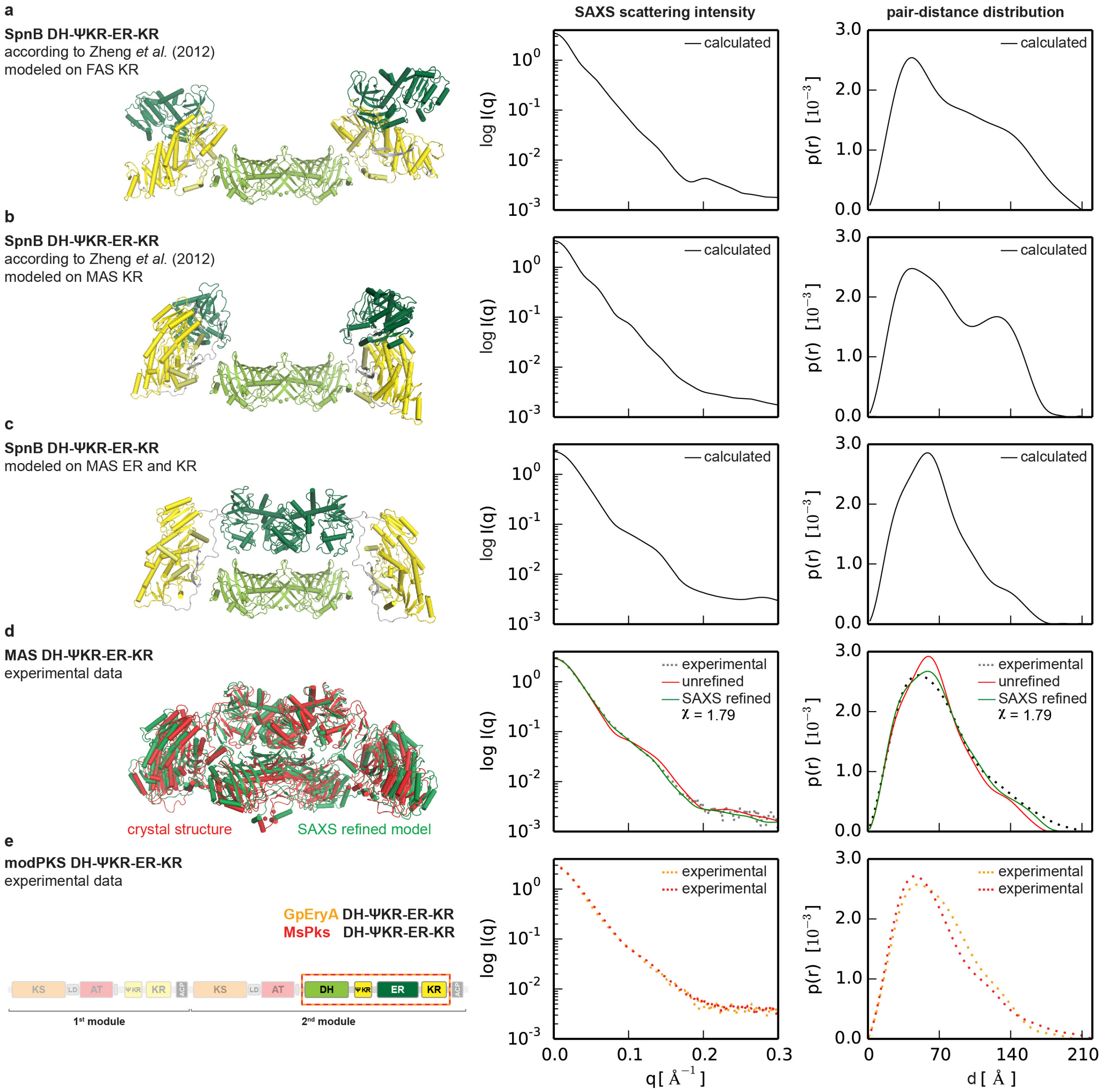
**Extended Data Figure 5 | Helical organization of central linking segments in MAS and modPKSs. a**, Assembly of the MAS central linking region from authentic crystal structures of the condensing and modifying regions. The two structures overlap in sequence by four residues (blue). **b**, Hybrid model based on the homology completed KS dimer and reconnected helical linkers. Ends of loops defined by the KS–AT crystal structure are indicated by black spheres. Disordered segments in the dimeric condensing region are reconstructed by multi-template homology modelling (gold); colour coding is as in **a**. **c**, **d**, Helix formation in sequence regions corresponding to central linkers are also observed in the isolated crystal structure of the modPKS DH domain of the fully reducing DEBS module 4 (ref. 16) (**c**), RifDH$_{10}$ (ref. 17) (not shown), and in the crystal structure of the RhiE KS-B di-domain[24] (**d**), where a KS domain is connected directly to a DH homologous domain, the B domain.



**Extended Data Figure 6 | Analysis of structural variability in the modifying and condensing regions of MAS and related multienzymes.** **a**–**d**, Analysis of interdomain conformational variability between the 18 protein chains in the MAS modifying region crystal structure. **a**, **b**, Variability of ER positioning relative to DH from two perspectives reveals a screw axis motion combining translation of up to 8.5 Å with rotation of up to 13.6°. **c**, **d**, Variability of ΨKR/KR domain orientation relative to DH (**c**) and ER (**d**), respectively, reveals a hinge located in the interdomain linker region. **e**, **f**, Top and front view of six overlayed KS–AT di-domain structures[6,11,12,66,67] as indicated and the derived rotational distance of AT positioning around a common hinge in the LD. **a**–**f**, Relative locations of individual structures are highlighted by representative coloured helices. Translational components are indicated with an arrow on the rotation axes with signs indicated on the principle axis (thick, coloured according to the moving domain). All structures are aligned to a MAS reference domain (coloured ribbon). Rotation axes are shown for rotations larger than 6° and arrows are shown for translations larger than 1 Å.



**Extended Data Figure 7 | A comprehensive phylogenetic analysis classifies MAS into the branch of modPKSs .** Phylogenetic trees for 55 fully reducing MASs/PKSs/FASs modules were constructed on the basis of only KS domains (**a**), complete condensing regions (**b**), the ER domain (**c**), or all catalytic domains (**d**). *M. smegmatis* MAS (MAS (Ms), bold, italic) and Msl-PKSs (italic) are more closely related to modPKSs (light green) and distinct from fiPKSs (blue) and animal FASs (yellow). All modules are labelled as protein name (organism abbreviation) Uniprot number. Units are given as amino-acid substitutions per site. Indices correspond to Extended Data Fig. 5.



**Extended Data Figure 8 | SAXS analysis supports a MAS-like organization of PKS modifying regions.** Models (left) of modifying region organization and their respective theoretical and experimental scattering curves as well as pair–distance distributions (right) are shown. **a, b**, As proposed in ref. 19, the intact SpnB modifying region was modelled on the basis of the domain-swapped SpnB ER–ΨKR/KR structure, using either the structure of FAS (**a**) or of the MAS modifying region (**b**) as a guide for positioning KR relative to DH. The SpnB DH structure was generated by homology modelling. **c**, Model of the intact SpnB modifying region with dimeric DH and ER based on the structure of the intact MAS modifying region. **d**, Crystal structure of MAS before and after fitting to experimental SAXS data. A good fit ($\chi = 1.79$) is obtained by fitting SAXS data with a single model corresponding to an average conformation of the MAS structure. **e**, Sequence organization of two authentic modPKS modifying regions of similar ER–KR linker length to SpnB (left), together with experimental SAXS scattering data (right). The data closely match calculated scattering curves for a MAS-like architecture, but disagree with models based on a monomeric ER as suggested for SpnB.

**Extended Data Table 1 | X-ray data collection and processing table**

**a**

| | **KS-AT<br>1-892** | **DH<br>884-1189** | **DH<br>884-1189** | **DH-ΨKR-ER-KR<br>884-2020** |
|---|---|---|---|---|
| **Data collection** | | | | |
| Space group | P $4_12_12$ | P$2_1$ | P$2_12_12$ | $P_1$ |
| Cell dimensions | | | | |
| a, b, c (Å) | 77.53, 77.53, 371.22 | 59.65, 162.40, 66.62 | 67.06, 162.20, 59.49 | 151.38, 190.37, 270.84 |
| α, β, γ (°) | 90.0, 90.0, 90.0 | 90.0, 91.4, 90.0 | 90.0, 90.0, 90.0 | 95.6, 91.9, 103.7 |
| Resolution (Å) | 92.81 - 2.20 | 66.65 – 1.75 | 47.99 - 1.45 | 78.62 – 3.75 |
| $R_{merge}$ (%)* | 8.7 (135.8) | 3.6 (123.2) | 4.3 (134.5) | 25.0 (315.9) |
| *I*/σ*I** | 17.98 (2.26) | 12.32 (1.13) | 19.07 (1.36) | 8.95 (1.00) |
| $CC_{1/2}$* | 99.9 (67.7) | 99.9 (46.2) | 100.0 (76.6) | 99.6 (44.9) |
| Completeness (%)* | 99.6 (99.2) | 96.8 (91.6) | 99.3 (97.0) | 99.0 (98.9) |
| Redundancy* | 12.9 (13.2) | 3.9 (3.8) | 6.5 (6.5) | 9.0 (9.3) |
| Unique reflections* | 58,436 (9,008) | 123,118 (8,582) | 114,826 (18,878) | 296,164 (21,835) |
| | | | | |
| **Refinement** | | | | |
| Protomers | 1 | 4 | 2 | 18 |
| Resolution (Å) | 54.82 - 2.20 | 66.6 – 1.75 | 47.97 - 1.45 | 78.62 – 3.75 |
| $R_{work}$/ $R_{free}$ | 0.21 / 0.23 | 0.18 / 0.20 | 0.15 / 0.18 | 0.23 / 0.24 |
| No. atoms | 12,279 | 18,419 | 9,282 | 262,498 |
| Protein | 11,984 | 17,273 | 8,725 | 260,724 |
| Ligand/ion | -- | 408 | 60 | 1774 |
| Water | 295 | 738 | 497 | -- |
| B-factors | 81.91 | 57.63 | 35.29 | 171.34 |
| Protein (Å$^2$) | 82.38 | 56.90 | 34.55 | 171.44 |
| Ligand/ion (Å$^2$) | -- | 76.57 | 52.12 | 157.33 |
| Water (Å$^2$) | 63.04 | 64.27 | 46.27 | -- |
| R.m.s deviations | | | | |
| Bond lengths (Å) | 0.003 | 0.010 | 0.011 | 0.008 |
| Bond angles (°) | 0.73 | 1.10 | 1.31 | 0.97 |

**b**

| | **MsMAS DH-ΨKR-ER-KR** | **MsMAS DH-ΨKR-ER-KR-ACP** | **MsPks DH-ΨKR-ER-KR** | **GpEryA DH-ΨKR-ER-KR** |
|---|---|---|---|---|
| Beamline | X12SA (SLS) | X12SA (SLS) | X12SA (SLS) | X12SA (SLS) |
| Wavelength (Å) | 1.00000 | 1.00000 | 1.00000 | 1.00000 |
| Detector distance (m) | 2.14 | 2.14 | 2.14 | 2.14 |
| q range (Å$^{-1}$) | 0.005-0.707 | 0.005-0.707 | 0.005-0.707 | 0.005-0.707 |
| Capillary diameter (mm) | 0.1 | 0.1 | 0.1 | 0.1 |
| Scan lengths / step size (mm) | 4.5 / 0.5 | 4.5 / 0.5 | 4.5 / 0.5 | 4.5 / 0.5 |
| Positions / acquisitions | 10 / 10 | 10 / 10 | 10 / 10 | 10 / 10 |
| Scan repeats | 8 | 8 | 8 | 8 |
| Exposure time (sec) | 0.04 | 0.04 | 0.04 | 0.04 |
| Concentration (mg ml$^{-1}$) | 3, 6 | 5, 10 | 6 | 3, 6 |
| Temperature (K) | 285 | 285 | 285 | 285 |
| I(0) (Å$^{-1}$) [from P(r)] | 3.00 ± 0.00 | 2.26 ± 0.00 | 2.08 ±0.00 | 2.00 ± 0.00 |
| Rg (Å) [from P(r)] | 58.25 ± 0.09 | 66.90 ± 0.19 | 55.05 ± 0.09 | 56.54 ± 0.09 |
| I(0) (Å$^{-1}$) [from Guinier] | 3.01 ± 0.01 | 2.19 ± 0.01 | 2.1 ± 0.01 | 1.98 ± 0.01 |
| Rg (Å) [from Guinier] | 57.2 ± 0.30 | 62.23 ± 0.63 | 54.7 ± 0.29 | 54.9 ± 0.23 |
| Dmax (Å) | 201 | 250 | 191 | 192 |
| Model fit (χ) | 1.79 | N/A | N/A | N/A |

**a**, Crystallographic data collection and refinement statistics. The resolution cutoff was determined by $CC_{1/2}$ criterion (ref. 39). **b**, SAXS data collection and processing.
*Highest-resolution shell is shown in parenthesis.

**Extended Data Table 2 | Structural comparison and interface analysis**

**a**

| Structure 1 | Structure 2 | C$_\alpha$ r.m.s.d. [Å] | Aligned residues |
|---|---|---|---|
| MAS KS | DEBS KS$_5$ | 1.17 | 348 |
| MAS KS | DEBS KS$_3$ | 1.21 | 349 |
| MAS AT | DEBS KS$_5$ | 1.77 | 294 |
| MAS AT | DEBS KS$_3$ | 1.21 | 287 |
| MAS DH | Rif DH$_{10}$* | 1.86 | 249 |
| MAS DH | CurF DH* | 1.61 | 259 |
| MAS DH | CurJ DH* | 1.88 | 252 |
| MAS DH | CurK DH | 1.68 | 257 |
| MAS DH | CurH DH | 1.93 | 258 |
| MAS DH | DEBS DH$_4$ | 1.82 | 244 |
| MAS DH | MAS DH P2$_1$2$_1$2 / P2$_1$ | 0.70 | 284 |
| MAS DH P2$_1$ | MAS DH P2$_1$2$_1$2 | 0.40 | 284 |
| MAS ER | SpnB ER | 1.98 | 298 |
| MAS ER$_{NB}$ | PpsC ER$_{NB}$† | 1.13 | 170 |
| MAS ER | JamJ ER* | 1.63 | 308 |
| MAS ER | CurF ER* | 1.72 | 309 |
| MAS ER | CurK ER | 2.27 | 306 |
| MAS ER | mFAS ER | 2.15 | 299 |
| MAS ΨKR/KR | Tyl ΨKR/KR$_{10}$* | 1.83 | 380 |

**b**

| Interface 1 | Interface 2 | Area [Å$^2$] | SDV [Å$^2$] | Min [Å$^2$] | Max [Å$^2$] |
|---|---|---|---|---|---|
| KS | KS‡ | 1,759 | | | |
| | post-AT linker | 770 | | | |
| LD | KS | 521 | | | |
| | AT | 236 | | | |
| DH (P2$_1$) | DH (P2$_1$) | 979 | 59 | 938 | 1,021 |
| DH (P2$_1$2$_1$2) | DH (P2$_1$2$_1$2) | 999 | | | |
| DH (P$_1$) | DH (P$_1$) | 961 | 13 | 943 | 978 |
| ER | ER | 1,424 | 20 | 1,397 | 1,449 |
| DH-ΨKR-ER-KR | DH-ΨKR-ER-KR | 2,688 | 112 | 2,547 | 2,904 |
| DH$^d$ | ER$^d$ | 532 | 105 | 345 | 638 |
| DH-ΨKR linker | ΨKR/KR | 1,117 | 24 | 1,077 | 1,172 |
| | ΨKR-ER linker | 695 | 9 | 672 | 711 |
| ΨKR-ER linker | ΨKR/KR | 975 | 28 | 926 | 1,027 |
| | ER-KR linker | 471 | 24 | 426 | 517 |
| | ER | 353 | 20 | 330 | 411 |
| ER-KR linker | ER | 432 | 24 | 399 | 480 |
| | KR | 547 | 14 | 527 | 574 |
| ΨKR-ER / ER-KR linkers | KR | 1,111 | 22 | 1,076 | 1,162 |
| | ΨKR/KR | 1,488 | 25 | 1,449 | 1,544 |
| | ER | 683 | 30 | 637 | 735 |

**a**, C$\alpha$ root mean squared deviations obtained for structural comparison of MAS domains with their closest structural neighbours. **b**, Interfaces in the crystal structures of MAS variants. Standard deviations (SDV) and minima/maxima are given for structures containing more than one interface; d, dimer.

*Not part of a fully reductive modifying region,

†Protein Data Bank accession number 1PQW (unpublished).

‡By direct superposition of the monomeric KS–AT structure on the DEBS KS5 dimer. In the KS–AT dimer with restored interface (by homology modelling), the total area increases to 2,289 Å$^2$.

CORRIGENDUM

doi:10.1038/nature16487

# Corrigendum: Gigantism and comparative life-history parameters of tyrannosaurid dinosaurs

Gregory M. Erickson, Peter J. Makovicky, Philip J. Currie, Mark A. Norell, Scott A. Yerby & Christopher A. Brochu

*Nature* **430,** 772–775 (2004); doi:10.1038/nature02699

Questions have been raised about the methods used and conclusions reached in this Letter[1]. In revisiting the work, we realized that we did not provide sufficient methodological details regarding the many steps that went into our growth curve analysis, although the main conclusions of the paper were not affected. We regret any misunderstanding that might have resulted. A detailed rationale is available in the Supplementary Methods and Discussion of this Corrigendum and the source data are provided as Supplementary Data. We thank N. Myhrvold for bringing these issues to our attention.

In our reanalysis we found a minor translational mistake affecting the reported growth for *Tyrannosaurus*, which does not appear to have contributed to Myhrvold's concerns (details can be found in the Supplementary Methods and Discussion to this Corrigendum.) The correct equation is $\mathrm{Mass} = (5{,}649/[1 + e^{-0.55(\mathrm{Age}-16.2)}]) + 5$. This produces a maximal growth rate of 758 $\mathrm{kg\,yr^{-1}}$ using points closely bounding the inflection point and 774 $\mathrm{kg\,yr^{-1}}$ using the instantaneous equation. The reported value was 767 $\mathrm{kg\,yr^{-1}}$. This slight discrepancy (see the corrected Fig. 2 in the Supplementary Methods and Discussion to this Corrigendum) does not compromise our conclusion that *Tyrannosaurus* primarily achieved gigantism through evolutionary acceleration.

**Supplementary Information** is available in the online version of the Corrigendum.

1. Myhrvold, N. P. Revisiting the estimation of dinosaur growth rates. *PLoS ONE* **8,** http://dx.doi.org/10.1371/journal.pone.0081917 (2013).

CORRIGENDUM

doi:10.1038/nature16488

# Corrigendum: Dinosaurian growth patterns and rapid avian growth rates

Gregory M. Erickson, Kristina Curry Rogers & Scott A. Yerby

*Nature* **412,** 429–433 (2001); doi:10.1038/35086558

Questions have been raised about the methods used in the construction of dinosaurian growth curves in this Letter[1]. These were caused by ambiguity with regard to how curve-fitting functions were utilized, and insufficient explanation for how maximum growth rates were calculated. Taken together, these omissions gave the impression that we were able to fit very specific curves even in cases where data were seemingly too scarce to justify them. We apologise for the confusion. However, the main conclusions of the paper were not affected. A detailed rationale is available in the Supplementary Methods and Supplementary Discussion of this Corrigendum and the source data are provided as Supplementary Data. We thank N. Myhrvold for bringing these issues to our attention.

In our reanalysis we found the following translational mistakes, which do not appear to have contributed to Myhrvold's concerns; however, we take this opportunity to rectify them. The growth rates for *Psittacosaurus mongoliensis* were incorrectly reported as 5.82 $\mathrm{kg\,yr^{-1}}$ versus 5.28 $\mathrm{kg\,yr^{-1}}$ in Fig. 2 and 12.5 $\mathrm{g\,d^{-1}}$ in the legend to Fig. 3. Fortunately, the correct value of 14.1 $\mathrm{g\,d^{-1}}$ was used in the comparative regression calculations. Finally, the mass estimate used for one of the *Apatosaurus* specimens was incorrectly transcribed. This modestly affected the growth curve parameters in Fig. 2. Details can be found in the Supplementary Methods and Discussion to this Corrigendum along with the corrected Fig. 2. The change causes a negligible shift in the overall dinosaur regression line slope (see the Supplementary Data to this Corrigendum) and does not compromise our conclusion that dinosaurs grew like endotherms.

**Supplementary Information** is available in the online version of the Corrigendum.

[1] Myhrvold, N. P. Revisiting the estimation of dinosaur growth rates. *PLoS ONE* **8,** http://dx.doi.org/10.1371/journal.pone.0081917 (2013).

**PUBLISHING** Young researchers consider preprints **p.541**

**JOB TRAINING** Online courses for postdocs **p.541**

**NATUREJOBS** For the latest career listings and advice **www.naturejobs.com**

PLANET FLEM

MANAGEMENT

# When jobs go wrong

*Having to dismiss lab members is not easy, but there are ways to make the process less painful for all involved.*

**BY CHRIS WOOLSTON**

Most principal investigators (PIs) are eager to talk about their management success stories: postdocs and graduate students who have gone on to become science rock stars. But there's another reality of science that is rarely discussed. Sometimes, the relationship between PI and junior scientist crumbles beyond repair, or funds that support the lab dry up — and the PI must let that person go.

These are delicate matters — so delicate that lab leaders are often reluctant to consider the option. But there's a good chance they'll eventually have to. "Most PIs will have to go through this," says Karen Peterson, director of the Office of Scientific Career Development at the Fred Hutchinson Cancer Research Center in Seattle, Washington. And if it happens early on, it can upset careers. "Young PIs have no idea what they're supposed to do," she says. "They usually make a mistake the first time around."

Whether they have to sack a graduate student or postdoc because of misconduct, poor performance or a funding shortfall, PIs must take care to handle the situation in the right way, experts say — emotionally, and in terms of policy and legal requirements. Messy dismissals can damage the reputations of PIs as well as the person losing their position, and PIs who terminate someone without following proper procedure may be opening the door to litigation. Young PIs should therefore understand and embrace the policies at their institutions, and familiarize themselves with legal issues relating to employment, before they go ahead. They also need to adopt an approach that causes the least amount of trauma for everyone involved. Both sides can survive a sacking, if it's done with care.

## TOUGH DECISIONS

A UK biologist was flummoxed when one of his graduate students kept tequila in a lab drawer, refused to take notes during a meeting and botched every procedure discussed ▶

▶ at that meeting. For the biologist — who didn't want to be named, a sign of the stigma that surrounds the issue — these misdeeds and others added up to grounds for dismissal. The student was very enthusiastic about the research topic, but that eagerness wasn't translating to productivity, the biologist says. As a new PI, the biologist couldn't afford that kind of drag on his lab. "He produced no usable data and used valuable equipment time," the biologist says.

The PI says that he gave the student many chances to improve his performance, including step-by-step instructions for living up to the expectations of the lab. When the student didn't follow those instructions, the PI documented every misstep — every spreadsheet that never got corrected, every experiment that didn't get done.

The PI called a special meeting of the student's thesis committee, at which the student was told that his performance wasn't acceptable. This was followed by a last-chance meeting a couple of months later. The student was still in his first year, which eased the difficulty of decision to let him go.

At this university, as with many others in the United Kingdom, students are supposed to show 'acceptable progress' before they can start their second year of study. "All new PIs are advised to take this progression very seriously," the biologist says. "It's the easiest time to deal with bad students."

## PROCEED WITH CARE

Unless a student or postdoc has committed an egregious violation of scientific ethics or workplace protocol — such as fabricating data or assaulting another lab member — the route to a potential termination should be travelled slowly, deliberately and with careful documentation, experts say. "You first have to have a conversation," Peterson says. "It's a verbal warning: 'Here's where you are, and here's where I need you to be'." That warning should include specific and measurable steps that the lab member needs to take to get up to speed. If those benchmarks aren't met, the PI should issue a written warning that again spells out the steps needed to meet expectations. If the situation doesn't improve, the PI should start going down the long path towards termination, a process that can vary depending on a team member's title and position.

> ***"Scientists tend to be so respectful of each other that they're not clear in their communication."***

Graduate students generally aren't considered employees, but they are still protected by the policies of the institution. PIs who decide to dismiss a graduate student generally have to get approval from the thesis committee or the departmental graduation programme, a process that becomes much more difficult once a student has passed qualifying exams or has been approved for a second year of study. Looking back, the UK biologist is glad that he had regular meetings with his student from the beginning, giving him a chance to spot trouble early on. "When the student didn't respond to help and advice, we had plenty of time to go through a formal process before the one-year deadline," he says.

Postdocs, by contrast, are generally employees of their institution, so they fall in a different category. In the United States, their job security lies almost entirely in their contract, says Stephanie Caffera, a partner with the global law firm Nixon Peabody in Rochester, New York. As she explains, if there's no contract in place, "you have no right to your job". Lab employees in the United Kingdom, however, are safeguarded by not just contractual rights, but also a law that prohibits unfair dismissal for anyone who has been employed in the same job for two years, says Jane Byford, a partner with the firm Veale Wasbrough Vizards in Birmingham, UK. "You have to have a fair reason for termination," she says. This can include a shortage of funds, a documented lack of performance or lab misconduct.

Many postdoc contracts, however, include an initial six-month probationary period, during which they can be dismissed relatively easily. After that, a PI must provide documentation to the institution's human-resources department to justify the move, and the postdoc will need ample time, which should be spelled out in the contract, to look for another position. At Fred Hutchinson, Peterson says, postdocs are entitled to a six-month warning before they are let go, unless they've done something serious enough to warrant a quick dismissal.

The consequences of a misstep in the termination process can be severe. In both the United Kingdom and the United States, postdocs can, and do, bring suits against universities for wrongful termination. As Byford explains, UK employees can appeal their terminations to a government employment tribunal, and if it's found that they have been unfairly terminated, the university or institution may have to pay a fine of up to a full year's salary. Lawsuits in the United States can be even more expensive to the institution. "Universities are great fodder for plaintiff lawyers," says Caffera. Although some states, including New York, allow employees to sue their PIs directly, in most cases the universities will be on the hook for any payouts. And although the PI might not face fines, the damage to his or her reputation could be substantial, she says.

In many countries, dismissed workers can potentially sue for discrimination if they feel that they were let go because of their gender, age, race or other non-work-related reasons. Caffera says that this scenario underscores the importance of documentation — thorough and careful records of infractions in the lab could someday prove crucial to the defence of a discrimination lawsuit.

### TAKE PRECAUTIONS

*How to fireproof your lab*

Lab lay-offs can damage the career and reputation of everyone involved, so young principal investigators (PIs) should minimize the risk that anyone will ever have to be let go, says Karen Peterson, director of the Office of Scientific Career Development at the Fred Hutchinson Cancer Research Center in Seattle, Washington. She encourages PIs to think about fireproofing their labs when interviewing prospective members. Above all, potential postdocs and graduate students need to know the rules and expectations of the lab, and the more specific the better. "They know going in what to expect, and they can choose not to work there," Peterson says. If a postdoc or student doesn't want to work with mice, for example, or for a PI who intends to monitor their work closely, some labs wouldn't be a good fit.

PIs should not get upset if prospective postdocs decide to work elsewhere, Peterson says. And if they find themselves complaining that there aren't enough good postdocs available, they may want to take a look at their approach. "Their expectations may be unrealistic," she says.

When conflict bubbles up, early intervention can help to keep the lab intact, Peterson says. "The PI should talk directly to that person in private," she emphasizes. The conversation should focus on specific missteps and their impacts on the lab. "Don't say, 'You're being a jerk,' and don't say something inflammatory," she cautions. "Just mention that they've been showing up late to lab meetings, and then shut up and listen. You might learn something that can be changed."

As an ombudsman for her institution, Peterson has seen first-hand how consultation with a neutral third party can often help to prevent a dismissal. "PIs come to me because they want to keep a person," she says. By giving both sides a chance to explain their issues, she can often find enough common ground to keep a lab together. **C.W.**

Even when there is no conflict, PIs may have to sack lab members when funds evaporate unexpectedly. Darren Boehning, a molecular biologist at the University of Texas in Houston, has twice had to reluctantly let go of postdocs when grant money dried up prematurely. In one case, the postdoc had only a month's notice. "Every postdoc contract I've seen says that the position is dependent on funding," he says. In this case, he knew of colleagues who were looking for a postdoc, and the individual was able to move to another lab. She eventually went on to a faculty position — as did the other postdoc who was released ahead of schedule. "You have to help them transition if you can," Boehning says. Not only can such support help to save the career of the person who is being let go, it can protect the PI's reputation.

### CLEAR COMMUNICATION

Graduate students and postdocs at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany, rarely leave their labs before the end of their contracts, says Helke Hillebrand, academic coordinator and dean of graduate studies. Once they pass their one-year probationary period, graduate students are under contract with the institution, which means that any dismissal would have to involve the human-resources department and the graduate-studies office. "They would never be totally dependent on their supervisors to determine their fate," she says.

As with other institutions in the United Kingdom and mainland Europe, EMBL requires graduate students to finish their degree within four years, a rule that puts pressure on everyone to keep student–mentor relationships intact. If a student has to change labs more than halfway through their training, it will be nearly impossible for them to finish in the allotted time, Hillebrand says. After putting so much investment in a student, the institution is highly motivated to mediate any disputes between students and their PIs. "Students are a precious resource for research, so this protects the PI as well as the student," she says.

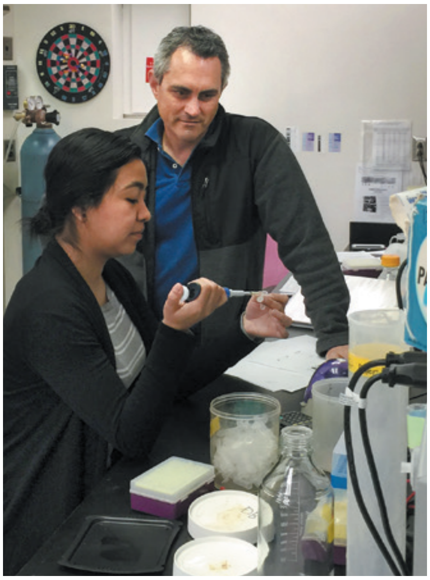
Geneticist Koen Venken has parted ways with three lab members since starting his lab in 2014 at the Baylor College of Medicine in Houston. When he first began to notice lax attitudes and poor production, he gathered the team for a PowerPoint presentation that spelled out his expectations. After seeing little progress, he repeated the presentation six months later. "They had plenty of time to identify their weaknesses and work on them," he says. He told the team that there wouldn't be a third PowerPoint warning. "I also indicated that I was more than happy to work with them to change for the better."

In retrospect, he sees that he might have avoided the dismissals had he been more up front about his standards before bringing anyone into the lab (see 'How to fireproof your lab'). He is now working on a formal agreement letter, complete with clearly stated expectations, that future lab members will have to sign before starting work.

When a PI does have to let a lab member go, it's important to keep the drama at a minimum by using a professional, straightforward approach, says Christopher Edwards, a science-career coach at Still Point Coaching and Consulting in Boston, Massachusetts, and the co-founder and former editor-in-chief of *Nature Biotechnology*. "There's a risk of having someone very angry with you after leaving your lab," he says. "One of my clients had to get a restraining order against a former grad student." He also knows of a case in which a disgruntled lab worker sued a former PI for plagiarism because the PI published a paper without including his name.

In Caffera's experience, messy break-ups can often be traced to a lack of clarity early on. "Scientists tend to be so respectful of each other that they're not clear in their communication," she says. "They speak obliquely. I would encourage them to be much more direct. People tend to assume they're doing a good job unless you tell them otherwise."

Laboratory lay-offs are likely to be far from the minds of most junior researchers — until they find themselves in a lab that isn't working. The silence around the issue makes it hard for PIs to anticipate or react to strife in their own labs. Venken hopes that other PIs can take something away from his experience. "It's very sensitive," he says. "But if no one is willing to talk about it, no one can learn from it." ■


SCOTT COLLUM

**Molecular biologist Darren Boehning works with graduate student M. Iveth Garcia.**

**Chris Woolston** *is a freelance writer in Billings, Montana.*

PUBLICATION

## Preprints pondered

A trio of commentaries explores whether it makes sense for early-career scientists to post public copies of articles before they are accepted by journals — or even submitted to them (see G. McDowell *F1000Research* **5,** 294; 2016). The authors, who include elite scientists, junior faculty members and postdoctoral researchers, examine whether depositing work on preprint servers is an opportunity or a vulnerability for young researchers. Early-career scientists harbour concerns about persuading colleagues to agree to a preprint, being ignored or receiving criticism on social media or from senior members of the field. But preprints also allow them to demonstrate their research productivity independently of unpredictable publishing timelines. It is unclear how preprints are taken into account by grant reviewers or hiring and promotion committees, and many researchers worry that the data could be used by rivals who might then beat them to publication. But early disclosure can also spark fruitful collaborations, says one author, who credits his preprint for initiating connections that accelerated his follow-up work. The commentaries are linked to last month's Accelerating Science and Publication in Biology meeting in Chevy Chase, Maryland.

TRAINING

## Postdocs to learn online

A group of prominent US scientists from the academic, government, industry and non-profit sectors aims to create an online training centre to collect career-development resources for postdoctoral researchers. Most postdocs end up in jobs away from the laboratory, but career-development training for them is patchy across institutions. The centre would be a repository for lesson plans, materials (including the individual development plan tool, a career-development workbook that is available online or through host institutions) and resources (such as a list of certified training advisers) to help universities to create career-development programmes. All such content on the website would be peer-reviewed and checked, and a steering committee will address specific issues, such as the target audience for lesson plans and how materials and career advisers will be vetted. The American Society for Biochemistry and Molecular Biology in Rockville, Maryland, has pledged to support the development of the centre with funding and staff time.

# THE HUMAN IS LATE TO FEED THE CAT

*A waiting game.*

BY BETH CATO

Sassafras was most displeased. The woman was late to return home. Sassafras positioned herself in the hallway with a clear vantage point of both the front door and the laundry room that contained her barren wet-food dish.

Her anxiety increased as the hours passed. She fully groomed herself five times over, and even lapsed her guard duties long enough to eat some dry kibble; that dish was despairingly low as well. Her snack done, she resumed her watch.

If Sassafras had to rise again, it would be for the curtains to know her wrath.

The front doorknob jostled. The cat hopped to her feet, fur like a puffy white cloud. The woman lurched inside and slammed the door behind her. Without sparing a glance at Sassafras — the nerve! — the woman dived through another doorway. With a dismayed glance towards the laundry room, the cat followed her.

The woman was ill. She radiated sourness, and the potency of it worsened as she leaned over the toilet. But Sassafras was the very model of support and patience. She purred and marched back and forth in the gap between woman and toilet, with brief pauses to groom the woman's elbows.

"No, Sassy." The woman pushed her away.

Confused, the cat scrambled backwards. The woman fell back on her haunches, groaning.

"People are coming down sick everywhere. Some virus. I heard the hospitals are full. I couldn't take the train. On the bridge, I saw someone …" She wiped her wrist against her mouth and shuddered. Sassafras smelled blood. "I'm afraid … I don't know … I tried calling my mom, but …" The electricity flickered like a pulse but stayed on.

Sassafras rubbed against the woman's knees and prodded her towards the doorway. The woman did move, but only to hover over the toilet again. The cat retreated a few steps with a despondent yowl.

"I know, Sass." The woman coughed and hacked. She pulled a towel from the rack and dragged it with her as she crawled down the hallway. Sassafras was unsure what to make of this sort of progression, but at least the human was going the right way. The cat trotted ahead to act as a guide, her tail like an exclamation point as they entered the laundry room.

ILLUSTRATION BY JACEY

The woman knocked the food box from the shelf and, after more delays, managed to rip open a pouch. The glorious perfume of savoury tuna in gravy filled the room. Sassafras purred like a motor and she settled in at her food bowl.

The lights flickered again. Outside, car horns blared, followed by pop, pop, pop, and a prolonged scream.

"Oh God. The world's gone to hell. What's going to happen, Sass?"

The cat felt the weight of the question and glanced up. She had just resumed eating when the sound of the jostled dry-food bag made her freeze. The woman had dragged the large bag from between the washer and dryer. She undid the clip at the top, then tipped the whole thing on its side. An avalanche of kibble tumbled across the floor.

Sassafras stopped eating her beloved tuna. Why was her crunchy food all over the floor? The woman knew Sassafras only ate dry kibble if it was in the appropriate bowl. The woman smelled increasingly wrong, too. Her body was too hot. It was rank. Sour. Unfamiliar. The cocoa butter lotion that the cat liked to lick from the woman's calves couldn't even be smelled now.

The woman used the supply shelf to pry her body upright. The effort left her wheezing and coughing. "At least it'll be fast for me, Sass. That's what the news was saying." Her laugh made her cough more. "To think, this morning … I thought it would be an awful day because I ran out of coffee."

The woman edged her way to the window. It took her several minutes of effort to crank open the pane of glass. Fresh night air flowed through the room. Sassafras's whiskers flared out as she breathed in the fragrance of trees and strange cats and city. She started forward but the woman wobbled and collapsed to her knees, forcing the cat back towards the doorway.

"I used to hate it when you hunted birds at our old place, remember? I would get so mad. But now …" She rolled to her side. Her shoulders racked as she coughed and choked. It took her a minute to speak again. "The tree branch goes right to the window. Good thing … landlord never had the landscapers come. You can go in and out, Sass. If I can get to the sink, I'll …"

More sickness, more coughing. Unsure of the strange assault of smells, Sass stayed back, ears flicking at the contrast of the woman's noises with the sounds from outside. The loudness there had frightened away the birds.

The woman's racket faded to weak sobs. That sound, the cat knew from nights when she shared her bed with the woman. Sass took mincing steps around the foulness on the floor and stopped at the woman's hands. The fingers twitched and managed to rest on Sassafras's sloped spine.

The cat lowered and folded her body into a bread-loaf form. The woman's hand grew heavier on Sass's back, and the cat purred. ■

**Beth Cato** *resides in Arizona. She is the author of the* Clockwork Dagger *steampunk fantasy series from Harper Voyager. Her website is BethCato.com.*